nature

14 June 2001 Volume 411 Issue no 6839

Towards new standards in university–industry collaboration

A new report provides a useful overview of issues surrounding the mutual and conflicting interests of academic researchers and business. Its recommendations should inspire caution rather than unalloyed enthusiasm.

"If you had time to patent it, it must be obsolete." That is one of the more memorable quotes from an important new survey of the issues arising from collaborations between US companies and universities. It reflects an ever-quickening world in which academic–industrial collaboration is the norm rather than, as some university administrators and company heads of research and development reportedly still see it, a questionable exception or even an aberration. On the contrary, 10% of all US academic funds now come from industry, and the money is growing.

The report was published last week by the Business–Higher Education Forum (see page 728), a partnership between the American Council on Education and the National Alliance of Business. It usefully highlights the need for universities to streamline their support and responsiveness in developing new collaborations — though not at the expense of quality control and protection of the university's core responsibilities. The information-technology (IT) sector especially, to which the above quote refers, is quick to exit when it encounters bureaucratic delays. The report is co-signed not only by the heads of top-quality research universities but also by the chief executives of major pharmaceutical and IT companies. It urges companies to give top-down backing for collaborations, so as to minimize obstacles from internal research groups protecting their turf.

The need for transparency and for the disclosure of conflicting financial interests and commitments have been called for often enough before, and receive due support here. Less often discussed are the troublesome issues of 'background rights', where a university is pressured to allow a collaborating company rights to technologies arising from university projects sponsored by others, including the federal government, in order to give the company the full value of the core inventions. Here the report aptly concedes that there is no easy answer. But it does not mention occasions when a company puts pressure on a university that can ill afford the heavy-duty legal work to explore the ramifications.

There are long-established concerns about publication delays while a company examines a discovery's commercial potential. And the report points to increasing pressures by industry to lengthen delays beyond three months. Valuably, it states that 60–90 days should be enough for most companies to express an interest and register for a patent — which, in the United States, leaves the way clear for publication.

The report is also helpful in pointing the way forward on research costs. It shows that there is no excuse for universities to succumb to pressures from companies and their own faculty to provide subsidized research to industry. It urges that universities charge federal rates for indirect costs.

In one key controversial area the report gives little help. It reiterates the well-known problems that arise when researchers develop materials and techniques that they want to patent or exploit, while the collaborating company may want to see those techniques widely used. But the report provides no recommendations.

There is also one major gap. The report says little about regulating and overseeing collaborations, or about public consultations concerning best practice. No one would suggest that there should be a central regulator — the main pressure for good behaviour comes from the conditions of funding set by agencies such as the National Institutes of Health, which can be powerful; but they cannot adequately protect the wider public interest in preserving universities' fundamental roles.

The report is welcome for what it is: the well-considered views of those with a vested interest in making such collaborations acceptable. It highlights occasions when companies have shown enlightened attitudes towards universities. And it represents good guidance. But it does not add up to a framework for resisting strong companies that take a predatory approach to academic collaboration. ■

An appropriate apology

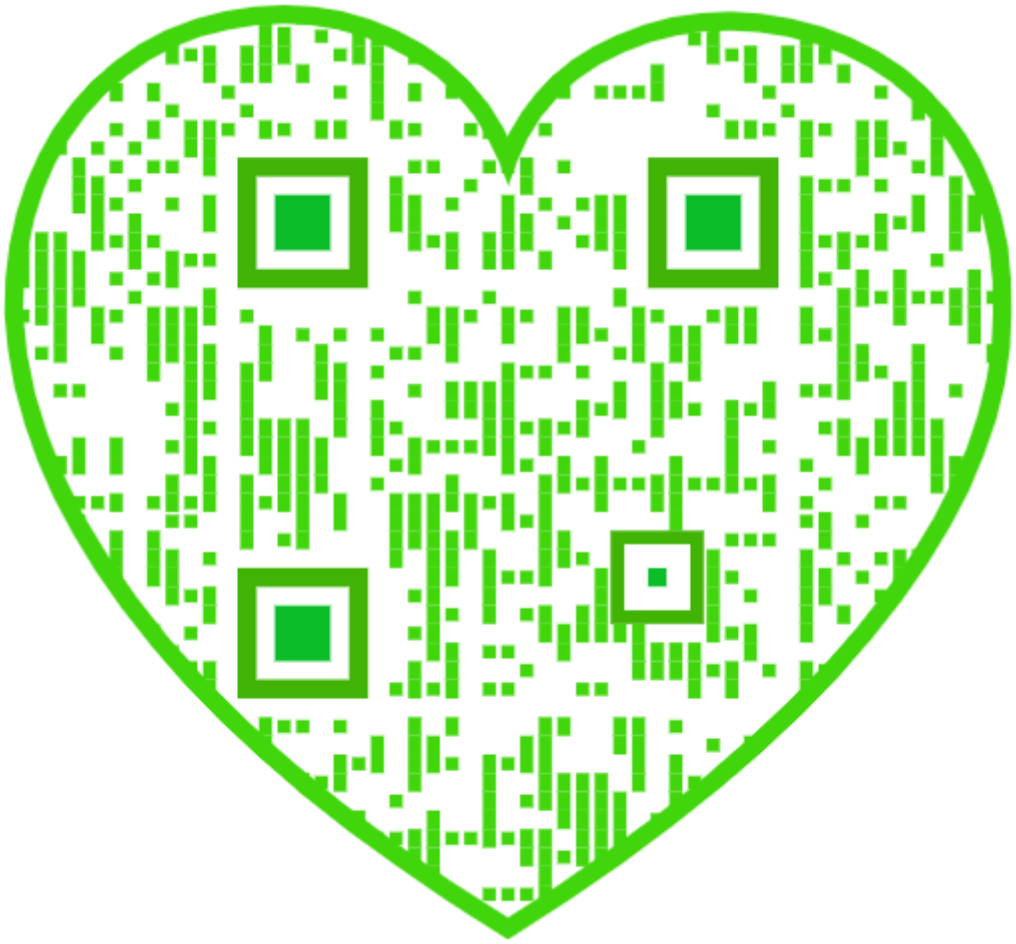
The president of the Max Planck Society has struck the right note in acknowledging past crimes against humanity.

It is wrong to dwell incessantly on the excesses of an increasingly distant past. Yet awareness of its past continues to haunt discussions in Germany about modern biology, particularly stem-cell techniques. And a history of barbaric science willingly carried out under tyranny may be all too relevant elsewhere today and in the future. So a statement last week by the head of Germany's most prestigious scientific institution deserves widespread attention.

Hubert Markl's thoughtful apology (see page 726) for the violations of human dignity perpetrated by scientists of the Kaiser Wilhelm Society, which was dissolved after the Second World War and re-established as the Max Planck Society (MPS), was well expressed. Avoiding self-serving rhetoric, he reflected that no one has the right to seek forgiveness, or alleviation of guilt, for "inextinguishable shame"; they can only express recognition of responsibility, and regret. In his speech to science historians and survivors of Josef Mengele's experiments in Auschwitz, Markl expressed a truth that Germans often flinch from articulating: that Germany was not the first and, tragically, not the last to develop an inhuman, racist regime that used genocide as a political instrument.

When his presidency of the MPS ends next year, Markl can take pride in his achievement of having opened up not just the archives of the MPS to independent historians, but also a better-informed debate on how German scientists can appropriately reflect on the Nazi past while embracing a confident future. ■

© 2001 Macmillan Magazines Ltd



Terror tests
German apology to victims of Nazi camp experiments
p726

Bug busters
US universities gear up to tackle virulent pathogens
p727

Outvoted
Referendum upsets Japan's nuclear power plans
p729

Return trip
NASA unveils plans to revisit Mercury
p731

Bush set to clash with European leaders over carbon emissions

Irwin Goodwin, Washington

On his first visit to Europe as president, George W. Bush will tell European leaders this week that he will not accept mandatory limits on carbon emissions. He remains unmoved on the issue despite fresh confirmation from the US National Academy of Sciences that such emissions are probably responsible for global warming.

On 11 June Bush announced that he would support a new research initiative "to study areas of uncertainty and identify priority areas where investment can make a difference". He also proposed "a joint venture with the European Union, Japan and others to develop state-of-the-art climate modelling".

But Bush sought to shift the blame from the United States, which has 5% of the world's population but produces about 25% of all carbon emissions. "Our approach must be based on global participation, including that of developing countries whose net greenhouse-gas emissions now exceed those in developed countries," he said.

The academy study was requested by the White House last month (see *Nature* **411,** 255; 2001), and was chaired by Ralph Cicerone, chancellor of the University of California, Irvine. It found that the conclusion of the Intergovernmental Panel on Climate Change (IPCC) — that the global warming over the past 50 years is probably the result of increases in greenhouse gases — "accurately reflects the current thinking of the scientific community".

But the academy said that the summaries of IPCC reports placed less emphasis on uncertainty than the reports themselves, and suggested changes to the IPCC process.

The White House has been trying to ease worries about the official US stand on global warming. After rejecting the Kyoto Protocol on climate change in March, Bush set up a cabinet task force on the issue, headed by Vice-President Richard Cheney and Bush's national security adviser, Condoleezza Rice.

Just before the academy released its report on 6 June, Cicerone and panel member Sherwood Rowland briefed White House and state officials on its contents. Cheney and Rice then met with environmentalists to sketch out their plans for addressing climate change.

MARK WILSON/GETTY IMAGES

Warming up: cabinet members look on as Bush confronts climate change ahead of his visit to Europe.

"They are weighing very heavily what to do at home before deciding what to do abroad," says Eileen Claussen, president of the Pew Center on Global Climate Change, who attended one of these meetings.

A Bush spokesman says the academy report provided "a basis of sound science on which decisions can be made", adding that "the president agrees action has to be taken".

But Don Ritter, a former Republican congressman who now heads the National Environmental Policy Institute in Washington, says that the academy report may prove embarrassing for Bush. He says: "The president probably never heard of Seneca, the Roman philosopher, who said: 'Do not ask for what you will wish you had not got'." ■

http://www.nas.edu

Labs seek share of NIH spending

Meredith Wadman, Washington

Senior advisers to the National Institutes of Health (NIH) have called for a billion dollars a year to be spent on construction and renovation of biomedical research labs.

The proposed investment would dwarf the $75 million that Congress earmarked for construction from the biomedical agency's 2001 budget of $20 billion. But advocates of the change say that a boost in such spending is needed if research infrastructure is to keep pace with the doubling of the overall NIH budget between 1999 and 2003.

The recommendation was delivered in a report to acting NIH director Ruth Kirschstein on 7 June by a group commissioned in 1999 by outgoing NIH director Harold Varmus.

The group says that existing construction funding cannot keep pace with recent large increases in NIH funding for researchers. "We wanted …s people [to] understand you can't do research in a garage," says William Brody, president of Johns Hopkins University, who chaired the group.

But NIH officials fear that Congress may respond to the request with money that the NIH could otherwise spend on grants to researchers. "If there was to be interest in putting more money in construction … then they might try to find it somewhere else [in the NIH]," notes Kirschstein. She deferred a decision on the group's findings until the committee meets again in December.

Academic medical centres in the United States customarily pay for new laboratories themselves, hoping to recoup some of their costs from government grants. ■

© 2001 Macmillan Magazines Ltd

Meeting hints at thaw in relations between genome rivals

Colin Macilwain, Washington

Genome assemblers from the private and public human genome sequencing projects finally met in peace at a workshop on 6 June, and found plenty of areas of common interest to discuss.

The much-heralded meeting — initially promised as part of the truce between the rival genome projects announced on 26 June 2000 by President Bill Clinton (see *Nature* 405, 983–984; 2000) — took place on neutral territory, at the Howard Hughes Medical Institute's headquarters near Washington.

"The goal of the meeting is to try and move forward and find out what common goals we have," explained Gene Myers, a mathematician who helped to devise the gene-assembly programme used by Celera of Rockville, Maryland, for its private project. Myers co-chaired the meeting with David Haussler, a computer scientist at the University of California, Santa Cruz, who had a prominent role in the public genome project, which is led in the United States by the National Institutes of Health.

Participants at the meeting, including gene-assembly and annotation experts from both projects, agreed that future sequencing projects of large genomes are likely to use a combination of the whole-genome shotgun approach pioneered by Celera and the clone-by-clone approach used by the public genome project.

There was agreement that the particular nature of an organism's genome, and the uses to which it might be put, would determine the appropriate strategy. "My feeling is that one approach to genome sequencing does not fit all," said Evan Eichler of Case Western Reserve University at Cleveland, Ohio. Eichler said he hoped the meeting would mark "a new era of détente" between the private and public projects.

Researchers using the genome data asked for more effort from both projects to build on the draft sequences and to deal with the large error rates being encountered. "Right now, with both the Celera and the public data sets, it is difficult to tell the difference between an interesting variant and a mistake in the data," said Sean Eddy of Washington University in St Louis, Missouri.

The meeting was attended by Francis Collins, director of the National Human Genome Research Institute, and other leaders of the genome effort, but not by Craig Venter, president of Celera. ■

Eva Kor (left) and her twin sister Miriam lead out survivors following the liberation of Auschwitz.

Max Planck Society admits to its predecessor's Nazi links

Alison Abbott, Munich

Hubert Markl, president of the Max Planck Society (MPS), has accepted that the management and staff of its predecessor society were involved in Nazi war atrocities, and has apologized to their victims.

His statement was in response to the findings of a group of science historians he commissioned in 1999 to investigate the role played by basic researchers of the Kaiser Wilhelm Society during the Second World War. This society, which was succeeded after the war by the MPS, was responsible for medical experiments undertaken at concentration camps.

Markl has long resisted issuing what he said would be an 'easy apology' (see *Nature* **403,** 813; 2000). But now, he says, the historians have produced hard evidence that proves "beyond the shadow of a doubt that directors and employees at Kaiser Wilhelm Institutes co-masterminded and sometimes even actively participated in the crimes of the Nazi regime".

His apology was delivered at an emotionally charged meeting held by the historians on 7 June in Berlin, at the halfway point of their five-year investigation.

Eva Mozes Kor, who with her twin sister Miriam was one of about 250 subjects who survived a programme of human experimentation orchestrated by the geneticist Josef Mengele at the Auschwitz concentration camp in Poland, told the meeting of her personal experiences.

Three times a week, Mengele's twins were walked to the Auschwitz main camp for experiments. "We had to sit naked in a room. Every part of our body was measured, poked and compared to charts and photographed. Every movement was noted. We were injected with germs and chemicals and they took a lot of blood from us," she said.

Kor said that her way of coping was to forgive the perpetrators. However, a second survivor, Jona Laks, said that the crimes should be remembered and never forgiven.

Kor explained how she developed a life-threatening fever after one of the injections, and feigned recovery in order to rejoin the experiments. "Would I have died, Mengele would have killed Miriam with an injection to the heart and would have done comparative autopsies on our bodies. This is the way most of the twins died."

Markl said that his apology was not a request for "removal of guilt", as the crimes were too heinous to allow such a release. By apologizing, both personally and on behalf of the MPS in proxy for the Kaiser Wilhelm Society, Markl said: "I am referring to the sincerest expression of deepest regret, compassion, and shame at the fact that crimes of this sort were committed, promoted, and not prevented within the ranks of German scientists."

He also apologized for the fact that the MPS had taken so long to begin its investigation, which he claimed was partly due to "a lack of willingness on the part of [some] inside and outside the Max Planck Society to face up to their historical responsibility".

Kor, who was one of ten survivors present at the meeting and is a founder of the survivors' group Candles, called the apology a "courageous gesture". ■

http://www.candles-museum.com

© 2001 Macmillan Magazines Ltd

Pathogen threat spurs research initiatives

Rex Dalton, San Diego

Interest in the risks posed by highly contagious fatal pathogens is prompting several US universities to set up special laboratories and research programmes to study emerging microorganisms and bioterrorism.

"There is a realization that emerging viruses are not going to go away," says virologist Michael Buchmeier of the Scripps Research Institute in La Jolla, California. The nation "has been complacent", he says. Now the pathogens "are staring us in the face".

Deadly microorganisms targeted for study include the Ebola, Marburg and Lassa fever viruses from Africa. There is also interest in other arenaviruses previously found in Latin America but now identified in California. West Nile virus and dengue fever viruses are also emerging as potential threats to public health. And there is growing official concern about bioterrorism involving agents such as anthrax and Bolivia's Machupo virus (see *Nature* **411,** 232–235; 2001).

Many researchers say the study of these pathogens has been hampered by a lack of academic facilities, as they must be housed in labs with a tough safety designation known as Biosafety Level 4 (BSL-4).

Research is at present concentrated at the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, and at a US Army base at Fort Detrick in Maryland. The National Institutes of Health (NIH) also has a small BSL-4 lab for antibiotic-resistant tuberculosis research in Bethesda, Maryland. And last year the Southwest Foundation for Biomedical Research in San Antonio, Texas, expanded its BSL-4 facilities. These had been restricted to a containment box with access through gloved arms, and now have a walk-in laboratory where scientists wear protective suits.

These facilities were previously the only BSL-4 laboratories in the United States. But in August the University of Texas Medical Branch (UTMB) at Galveston will start building a $10.5-million facility that should be finished by late next year. Last December, the UTMB recruited C. J. Peters, the CDC's head of the Special Pathogens Branch, to run the BSL-4 lab. And last month, the NIH certified the university's safety plans.

The California Department of Health Services and the University of California, Davis, are meeting regularly to plan another facility, which they hope to build for an estimated $50 million. Current discussions envisage that the state will fund construction of the BSL-4 laboratory on the Davis campus, with the university operating the lab.

At Texas Tech University in Lubbock, meanwhile, officials are planning a $98.5-million research facility that would include a BSL-4 lab at the decommissioned Reese Air Force base outside Lubbock. Last month, Texas Tech's board of regents voted to seek $4 million for design plans for the facility, which would focus on bioterrorism.

And the University of New Mexico in Albuquerque is exploring funding possibilities for a BSL-4 lab to extend its research on hantaviruses, says biologist Terry Yates, the university's new vice-president for research.

Both Texas Tech and New Mexico may look to Congress for funding for their facilities, officials say. In 1995, Texas A and M at College Station failed to secure funds from Congress for a BSL-4 lab. Officials there say they remain interested, but are not actively planning a facility. But attempts by other universities to get congressional funding may reinvigorate Texas A and M's plans. ■

GENO LORO

Germ warfare: several labs dedicated to the study of emerging virulent pathogens are planned in the US.

UK election sees revamp for farming and environment

David Adam, London

Discredited by the bovine spongiform encephalopathy (BSE) epidemic and still reeling from the effects of foot-and-mouth disease, Britain's Ministry of Agriculture, Fisheries and Food (MAFF) has finally been put out of its misery.

In a government reorganization that quickly followed the Labour party's overwhelming re-election on 7 June, Prime Minister Tony Blair announced that MAFF will be replaced by a new department that will seek an integrated approach to managing agriculture and the environment.

The new Department for Environment, Food and Rural Affairs (DEFRA) will now coordinate and fund research previously supported by MAFF, as well as work concerned with the environment, rural development, countryside, wildlife and sustainable development. These additional functions are inherited from the former Department of the Environment, Transport and the Regions, which has been broken up. DEFRA will be headed by Margaret Beckett, a former head of the Department of Trade and Industry (DTI).

The rest of Britain's apparatus for managing research and science policy will remain largely intact, after an election campaign in which science-related issues played little visible role. Lord Sainsbury, who as science minister is responsible for day-to-day science policy and oversees the research councils that provide the main support for academic research in Britain, was expected to stay in the same job. Patricia Hewitt, formerly the minister responsible for electronic commerce, will replace Stephen Byers as head of the DTI, giving her Cabinet-level responsibility for science and technology.

David Shannon, chief scientist with the former MAFF, describes the plans for DEFRA as "ambitious". He says its creation was necessary to resolve conflicts in the way agricultural and environmental issues were handled.

But some point out that this joint responsibility could cause the new body problems, and question whether the changes have been adequately thought out. "I think there is a serious danger that this is just window-dressing," says a senior official at one scientific society. ■

© 2001 Macmillan Magazines Ltd

Fears rise over BSE infection in UK abattoirs

David Adam, London

British abattoirs that cull the cattle most at risk from bovine spongiform encephalopathy (BSE) could be spreading the disease to meat sold for human consumption, a report from the Royal Society and the Academy of Medical Sciences has warned.

The report, *Transmissible Spongiform Encephalopathies: Statement by the Royal Society and the Academy of Medical Sciences*, was issued on 5 June. It suggests that the sheer resilience of prion proteins could be undermining the government's strenuous attempts to keep BSE out of the human food chain.

Eight of Britain's 394 slaughterhouses are regularly used in the government's programme to cull all cattle over the age of 30 months on the basis that these animals pose the greatest BSE risk. The corpses of these older cattle are incinerated — but the same slaughterhouse facilities are also used to prepare meat from younger cattle destined for butchers and supermarkets.

These activities must be carried out on separate days, but the report says that this may not prevent meat intended for human consumption from becoming infected with prion proteins. These are the proteins that are thought to cause BSE in cattle and variant Creutzfeldt–Jakob disease (vCJD) in humans.

Brian Heap, vice-president of the Royal Society, says that results published last year showing that prions may still be active even after they have been heated to 600 °C (see *Proc. Natl Acad. Sci. USA* **97,** 3418–3421; 2000) throw doubt on the effectiveness of the routine sterilization carried out in slaughterhouses. "Although this observation needs to be confirmed, we need to establish conclusively that work surfaces and equipment in abattoirs are not contaminated after the usual cleaning and sterilization procedures," Heap says.

Contamination of meat with prions from infected brain tissue by butchers' equipment was recently blamed for a cluster of cases of vCJD in the Leicestershire village of Queniborough (see *Nature* **410,** 502; 2001).

The report also recommends that new methods be considered for disposing of the vast mountains of potentially contaminated animal matter being stored in the United Kingdom. Some 430,000 tonnes of meat and bone-meal and 200,000 tonnes of tallow are held in storage, awaiting disposal. "We believe the authorities should consider eliminating this waste with meat-eating bacteria or through incineration without air at 850 °C," Heap says. He also points out that storing the waste is expensive and potentially hazardous.

The new report follows close on the heels of a modelling survey, commissioned by the Department of Health to address concerns about the transmission of vCJD by surgical instruments. This report concluded that surgical transmission could increase the size of the epidemic by up to 10%.

In response to such fears, the government earlier this year launched a £200-million (US$280-million) programme to improve sterilization procedures, and to use disposable instruments for operations that carry a particularly high risk of passing on the infection. But this is unlikely to have eliminated the risk entirely.

The Royal Society/Academy of Medical Sciences report also highlights the need to develop sensitive and cheap tests that would allow large numbers of live animals to be tested for BSE infection before they show symptoms of disease. Current tests have only been validated on brain tissue from animals that are known to have BSE. However, researchers are now working to develop methods that might detect prions in tissues such as blood before the appearance of symptoms (see page 810).

Another aspect of the problem that is highlighted by the new report concerns the uncertain job status of many BSE researchers. The report warns that short-term contracts for postdoctoral researchers are hampering research because they make it difficult for young scientists to commit themselves to the field. Research agencies "should consider creating some prestigious longer-term (five- to ten-year) fellowships and longer-term grants", it says. ■

http://www.royalsoc.ac.uk/royalsoc/new_fr.htm

http://www.doh.gov.uk/cjd/riskassessmentsi.htm

Contracts offer conflict resolution

Mark Schrope, Washington

Firm contracts are critical to reducing conflicts in research partnerships between US universities and industry, according to a report prepared by senior representatives of both groups.

The study identifies conflicts of interest that arise when professors and universities have a financial stake in research results as one of the most troublesome problems. Individual researchers should consider stock divestment, full disclosure in all publications, or even resignation from particular projects, it says, whereas universities should develop multiple partnerships to avoid reliance on any one company.

The report, *Working Together, Creating Knowledge: The University–Industry Research Collaborative Initiative*, was released on 6 June. It is the result of a collaboration between the American Council on Education, an association of 1,800 universities and colleges, and the National Alliance of Business, an industry group.

"We have two different cultures that have to learn to work together," says Nils Hasselmo, co-chair of the panel that produced the report and president of the Association of American Universities, which represents the largest US research universities. The other co-chair was Hank McKinnell, chief executive of Pfizer, the New York-based pharmaceutical company.

McKinnell says that one of the most important recommendations is that collaborations should have contracts that cover every aspect from the outset. Many universities and researchers start working together with significant portions of these agreements incomplete, he says.

The report also advises companies that they should pay the full overhead rate set by the federal government for each university, over and above the direct costs of research. Traditionally, many companies have negotiated lower rates.

But Sanford Chodosh, president of Public Responsibility in Medicine and Research, an ethics organization based in Boston, says that the report does not go far enough in addressing the problem of conflicts of interest in clinical research. He says that the report failed to address, for example, the ways in which companies influence the choice of research results that are made public. "A lot of the stuff that gets done never gets published," he says. ■

© 2001 Macmillan Magazines Ltd

Referendum stalls Japanese nuclear power strategy

David Cyranoski, Tokyo

Japan's plans to adapt its nuclear power programme to burn recycled nuclear fuel have been sent reeling by a referendum in a small town on the country's west coast.

On 27 May, the residents of Kariwa in Niigata — a community of just over 5,000 people — rejected by 1,925 votes to 1,533 the introduction of mixed oxide (MOX) fuel at the nearby Kashiwazaki-Kariwa nuclear power plant.

MOX fuel mixes recycled plutonium oxide, extracted from spent nuclear fuel, with uranium oxide. Under Japan's plan, one-third of the country's 51 light water-cooled nuclear power stations would start to burn MOX by 2010.

Although the Kariwa referendum is not legally binding, government officials, including Prime Minister Jun'ichiro Koizumi, have said they recognize the need to get the backing of local people before proceeding with the plans.

Critics of nuclear-fuel reprocessing say that the vote could have global ramifications, by dimming the prospects that Japan will become a major user of the technology. "The Kariwa referendum could be the mouse that roared," says Paul Leventhal of the Washington-based Nuclear Control Institute.

But both industry and government officials said after the referendum that the plan to burn MOX fuel at the power station will still proceed, albeit after some delay.

During the Kariwa referendum campaign, opponents of the plan claimed that MOX was more radioactive than existing fuel. They also said that because MOX contains plutonium, it could be adapted for use in nuclear weapons. "The Japanese government has never levelled with the Japanese people about these issues," says Leventhal.

In addition, critics claim that there is no need to reprocess spent fuel into MOX, given the plentiful supply of cheap uranium, and that reprocessing is being pursued largely as a means of deferring the problem of disposing of nuclear fuel. "As soon at they throw out the idea of recycling the spent fuel, they have to say what they would do with it," says Aileen Smith of Green Action, a Kyoto-based environmental group.

"If the MOX fuel cannot be used at the reactors, we will have no place to put it," admits one government official responsible for nuclear-energy policy. The official adds that there was "absolutely no chance" that Japan would pay Russia to store spent nuclear fuel (see *Nature* **411,** 401; 2001).

The idea of storing spent fuel rather than reprocessing it is taboo in government and industry circles, as there are no plans to build a long-term storage repository. Japan is instead investing more than ¥2 trillion (US$17 billion) in a reprocessing plant being built at Rokkasho, Aomori.

O. SEMBA/AP

Voted out: locals have vetoed the use of recycled fuel at the Kashiwazaki-Kariwa nuclear plant.

Electric companies that plan to burn MOX in Japan can only do so with the agreement of the regional governor, and such agreements have been reached with the governors of Fukui, Fukushima and Niigata.

But the discovery of falsified inspection data for a shipment of MOX fuel from Britain to Fukui put that region's project on hold. That scandal, together with accidents such as that at Tokai in 1999 (see *Nature* **401,** 736; 1999), led the governor of Fukushima to ask for a year's delay on its MOX initiative. Now the Kariwa referendum looks set to delay the introduction of MOX in Niigata by a similar length of time.

Government and industry officials declined to say what would happen if the introduction of MOX is opposed at other power stations. One said that supporters of the plan will now embark on a programme aimed at "gaining the understanding of the people" concerning the safety of plutonium.

The United States abandoned nuclear-fuel reprocessing in the 1970s, but other countries have continued to pursue it, and MOX fuel is used to generate power in several countries, including France, Germany and India. In Japan, it has been burned in research reactors, but not in power plants. ■

Airborne telescope delayed as plane is made ready

Tony Reichhardt, Washington

The first flight of a long-awaited airborne telescope developed by German and US scientists will take place two years later than planned, NASA officials have admitted. Take-off is now scheduled for late 2004 or early 2005.

The Stratospheric Observatory for Infrared Astronomy (SOFIA) was conceived as a follow-on to the Kuiper Airborne Observatory, which ended a productive 20-year career in 1995. The $500-million project will use a German-built telescope the same size as the Hubble space telescope (2.5 metres in diameter) mounted in a converted Boeing 747.

The aircraft will fly three or four times a week from NASA's Ames Research Center in California, with a planned lifetime of 20 years. Cruising at altitudes of around 12,500 metres, above almost all of the infrared-absorbing water vapour in the atmosphere, SOFIA will study star and planet formation as well as the interstellar medium.

But modifying the airplane, which can stay aloft longer than a conventional 747, has needed more manpower and money than the company doing the job — Raytheon Aircraft Integration Systems in Waco, Texas — had bargained for. Faced with projected cost overruns of tens of millions of dollars, NASA has decided to delay the first flight by two years.

"We're not particularly happy about it," says Ken Ledbetter, director of the Flight Programs Division at NASA's Office of Space Science. But for an agency that is struggling to complete a series of missions started in the 1990s, it was the best available option. NASA science chief Ed Weiler even told a recent meeting of the National Research Council's Space Science Board that, were it not for the German involvement, he would have considered cancelling the project. ■

http://sofia.arc.nasa.gov

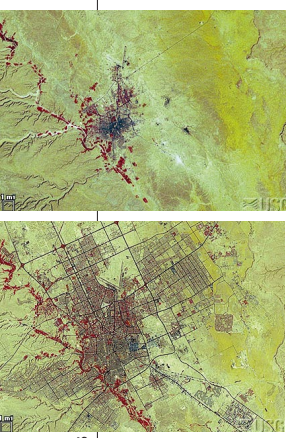
NASA

Alterations to the plane set to carry NASA's SOFIA telescope are holding up the project.

© 2001 Macmillan Magazines Ltd

United Nations gives global ecosystems a health check

Nairobi A four-year global study to assess the health of the world's ecosystems was launched last week by the United Nations Environment Programme (UNEP). The $21-million Millennium Ecosystem Assessment, a collaboration of some 1,500 scientists, aims to provide decision-makers with information about what can be done to reverse negative environmental trends.



USGS

Overview: satellite images show how ecosystems change.

The initiative builds on a pilot study published last year by the World Resources Institute, a Washington-based environmental group, which found that changes to the world's ecosystems are threatening biodiversity and human health, and increasing vulnerability to environmental disasters.

One of the first tasks for the assessment will be the analysis of NASA satellite images, which demonstrate recent changes in global ecosystems. "The satellite data will allow us to identify areas where direct scientific assessments by people on the ground are urgently needed," says Dan Claasen, head of UNEP's environmental information networking unit.

www.ma-secretariat.org

Computer scientist seeks right to reveal music notes

Washington A Princeton University computer scientist is going to court to secure the right to present his work at a conference in August.

Edward Felten cracked digital music anti-piracy technologies last year as part of a public contest sponsored by the recording industry. Felten had planned to present his work at a conference in April, but was warned in a letter from the Recording Industry Association of America that public disclosure of information gained from the contest could violate the terms of the competition and the 1998 Digital Millennium Copyright Act (DMCA).

Felten is now asking a federal court to rule that he is entitled to present his work at the USENIX Security Symposium in Washington. He is also challenging the DMCA, which blocks public access to technology that circumvents copyright-protection systems. Felten argues that, when used to prevent academic publication, the DMCA violates his constitutional right to freedom of speech.

Science park makes Québec a Biotech City

Montreal Plans to turn Québec into an international centre for science and technology took a step forward last week with the creation of Biotech City, a science and technology park in Laval, close to Montreal.

The Biotech City will receive Can$250 million (US$165 million) from Québec's provincial government over the next five years to promote the commercialization of life-sciences research and collaborative projects between universities and industry. The park will be based around the existing Armand-Frappier campus of the National Institute for Scientific Research (INRS), but several new institutes will also be created.

The creation of Biotech City is viewed by some as part of a plan by Québec's premier Bernard Landry to make the province independent from Canada. Since becoming premier in March, Landry has made the promotion of science and technology part of his strategy for independence.

www.citebiotech.com

© 2001 Macmillan Magazines Ltd

Irish vote hits prospects for Eastern European science

London Hopes that science in Eastern European countries would be boosted by the enlargement of the European Union (EU) suffered a setback on 7 June when, in a referendum, Ireland voted against the Treaty of Nice, which would have allowed for the expansion of the EU.

Scientists in the candidate countries in Eastern Europe are already eligible for funding from the EU, but the countries cannot apply for money for new research facilities. Scientific bodies in Eastern Europe hope that, once their countries join the EU, they will be able to apply for money from its structural funds, designed to help build infrastructure in Europe's poorer regions.

Enthusiasm for enlargement is now waning in several EU countries, including France, Spain and Portugal, prompting fears in the candidate countries that they may never gain access to the structural funds.

NASA sets sights on mission to Mercury

Washington NASA has approved a new $256-million mission to Mercury, scheduled to launch in 2004. The Mercury Surface, Space Environment, Geochemistry and Ranging (MESSENGER) spacecraft will fly by Venus and Mercury twice before beginning a one-year orbit of Mercury in 2009.

Mercury rising: NASA sets date for return trip.

M. ROBINSON, NORTHWESTERN UNIV.

The mission aims to collect information about the planet's crust, atmosphere and magnetic field. The spacecraft will be built at the Johns Hopkins University Applied Physics Laboratory in Laurel, Maryland. Preliminary work on the project began 18 months ago.

MESSENGER will be the second spacecraft to visit Mercury, after Mariner 10, which photographed less than half of the planet's surface in the 1970s.

http://sd-www.jhuapl.edu/MESSENGER

Inspections speed plans to reopen fast-breeder reactor

Tokyo Local government officials last week gave the go-ahead for inspections that could lead to the reopening of Monju, Japan's fast-breeder nuclear reactor.

Situated in Fukui province, 500 kilometres southwest of Tokyo, Monju has been closed since sodium coolant leaked from a cracked pipe and burst into flames in 1995 . No one was hurt, but local residents were angered by attempts to cover up the incident. The safety inspections, which will take one year to complete, will need to be followed by local government approval if the reactor is to begin operating again.

Groups campaigning against Monju have argued that plans to reopen the reactor reflect a desire by officials involved with the project not to lose face. The reactor is currently receiving over ¥8 billion (US$66 million) a year from the government to cover its maintenance costs.

http://www.jnc.go.jp/zmonju/mjweb

European council to advise policy-makers

London Plans for a new independent body to provide European Union (EU) policy-makers with scientific advice were announced this week by the national science academies of the EU member states.

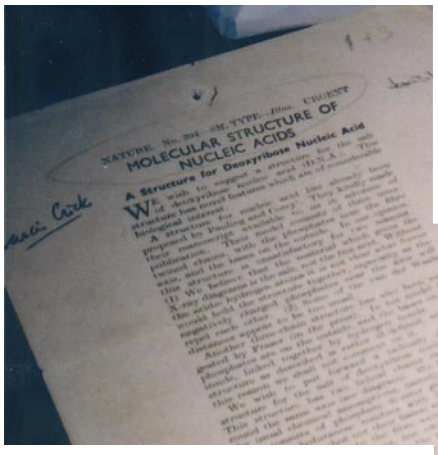
The European Academies' Science Advisory Council will provide advice on issues such as the environment, energy and food safety. The council, which will work with European scientists when preparing scientific advice, will be based at the offices of the Royal Society in London.

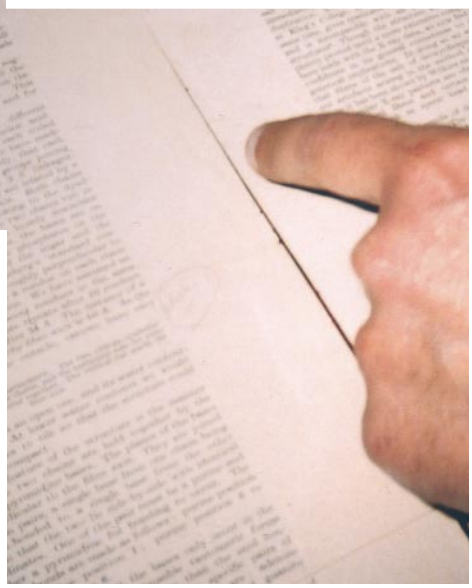
© 2001 Macmillan Magazines Ltd

The history man

A Californian book dealer has created a unique archive of molecular biology's achievements. But some historians are uneasy about such a valuable resource resting in private hands, says Rex Dalton.

REX DALTON





No. 4356 April 25, 1953 NATURE 737

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

Treasure trove: Jeremy Norman (top) displays a proof of Watson and Crick's DNA paper.

Collecting historic scientific documents seems to be in Jeremy Norman's genes. His late father, a prominent San Francisco psychiatrist, spent a lifetime building a library of medical books and records. Now the son — a dealer in rare books for more than 30 years — has created a unique collection: an archive of documents and other artefacts relating to the history of molecular biology.

Over the past four years, Norman has quietly been acquiring laboratory notebooks, correspondence, photographs and galley proofs of landmark journal articles. His collection is a treasure trove of material from many of the world's best-known biologists, including numerous Nobel laureates. Norman promises to make his archive available to science historians. But those assurances have not placated some researchers for whom the idea of such a resource being privately owned, rather than residing in a public museum, is anathema.

For anyone familiar with the early days of molecular biology, being shown the Jeremy Norman Molecular Biology Archive is a jaw-dropping experience. The collection is housed for now in a secure but secret location, until a permanent home is constructed in the San Francisco suburb of San Anselmo. Bespectacled and scholarly, Norman carefully pulls binders of records from a safe the size of a double-width refrigerator. The names on the files read like a *Who's Who* of the field's founders: Francis Crick, James Watson, Max Delbrück, Maurice Wilkins, Sydney Brenner and many more.

Controversial depiction

The file marked "Crick", for instance, contains the galley proof of the 1953 *Nature* article, co-authored with Watson, reporting the discovery of the double-helix structure of DNA. Norman acquired it from a third party, who obtained it from Crick years ago.

Norman also possesses one of the dozen or so original copies of *Honest Jim*, Watson's first draft of a manuscript that — in watered-down form — became *The Double Helix*. Even the published book was controversial, because of its colourful depiction of the players in the race to determine the structure of DNA — Watson famously said that he had never seen Crick "in a modest mood".

Another binder includes page after page of correspondence between Rosalind Franklin and her scientific colleagues. Watson and Crick, working at the University of Cambridge's Cavendish Laboratory, based their insights on Franklin's X-ray crystallography data, collected while working in Wilkins's lab at King's College London — and many scientists feel that her contribution was for years downplayed. "One letter is generally considered important," enthuses Norman. "But to have almost all the correspondence between two people is very rare."

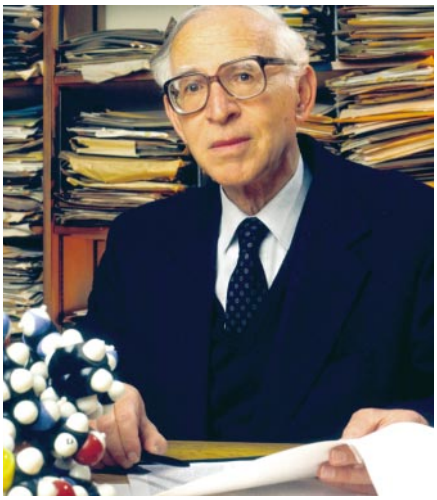
As Norman moves from file to file, the joy he takes from the collection is obvious. He has the air of a patron of the arts, and has been collecting rare books and documents since childhood. Shortly after graduating with a degree in history, he opened a shop selling rare books, specializing in science. Now 55, Norman has over the years bought and sold several collections of scientific and technical documents, always seeking to be, as he puts it, "ahead of the market". His current holdings include an archive relating to the history of aerospace, up to the 1957 launch of Sputnik 1, and another on the history of computing, from the seventeenth century until 1976.

But Norman insists that the molecular biology archive will never be traded. Eventually, he says, it will be donated to a public institution — possibly his *alma mater*, the University of California, Berkeley. In the meantime, he says historians and other researchers can seek access to the records.

Nevertheless, some science historians complain that there are no long-term guarantees, and fear that the collection may eventually be broken up and sold, with important papers disappearing into the drawing-rooms of wealthy collectors. Indeed, the

© 2001 Macmillan Magazines Ltd

J. KING-HOLMES/SPL



Handed over: Aaron Klug, 1982 chemistry Nobel laureate, sold his papers to the archive.

medical collection of Norman's father was auctioned piece by piece by Christie's in 1998. "Archives are for ever," argues Spencer Weart, director of the American Institute of Physics Center for History of Physics in College Park, Maryland. "You need an institution, not a private individual, in control."

Donation dilemmas

Norman rejects this criticism. "Before I did this, the papers of all these famous people were ignored; they were just sitting there," he says. "Now people are screaming. Had institutions preserved the papers, I wouldn't have been able to buy them."

Some experts have a degree of sympathy with this argument. Marcel Caya, an archivist at the University of Quebec in Montreal, is an advocate of public archives, but concedes that universities and libraries are often strapped for cash to catalogue and house scientists' donated papers. Wilkins, who shared the 1962 medicine Nobel with Watson and Crick, recalls trekking to the Churchill Library in Cambridge in the mid-1980s to review the papers of his former supervisor, John Randall. "They weren't organized," he says. "There was just piles of stuff in cardboard boxes." Eventually, says Wilkins, the library was "prodded" to catalogue Randall's papers.

Irrespective of fears about Norman's collection eventually being broken up, some archivists in Europe have expressed misgivings about scientific documents leaving the countries in which the work was done. Last year, for instance, the chairman of Britain's Historical Manuscripts Commission, Lord Bingham, wrote to Aaron Klug, then president of the Royal Society, noting his concern about the loss to British heritage represented by the export of the papers of pioneering British molecular biologists.

That letter placed Klug — who won the chemistry Nobel in 1982 for his development of crystallographic electron microscopy and studies of the complexes formed between nucleic acids and proteins — in a somewhat delicate position. In 1973, the Royal Society and the Historical Manuscripts Commission created a unit to catalogue the archives of contemporary scientists. Now based at the University of Bath, the unit encourages the donation of papers to scientists' own institutions, or other public archives. Yet Klug had sold his original papers to the Norman archive, rather than donating them to his institute, the Medical Research Council's Laboratory of Molecular Biology (LMB) in Cambridge.

Klug told *Nature* that his reply to the commission noted that he "had been given assurances that the collection is intended to finally be donated to a major academic institution in the United States where they would be securely kept and free access be given to researchers".

This seems to have satisfied the Historical Manuscripts Commission, but the idea of scientists selling their papers is still viewed with distaste by some researchers. There are rumours of large sums of money changing hands, but scientists whose papers are in the Norman archive are unwilling to discuss the issue. Klug and Wilkins declined to comment on the fees they had received when questioned by *Nature*, as did Max Perutz, also of the LMB, who shared the 1962 chemistry Nobel for his studies of protein structures.

Norman's agent in these deals was Al Seckel, a cognitive neuroscientist at the California Institute of Technology in Pasadena who trades in rare scientific documents in his spare time. Seckel appears to have turned up, offered to purchase documents for the Norman archive, and if a deal was struck, provided the scientists with copies of the originals to retain for their own files.

Free access

Other scientists have provided Seckel with material for free. Vittorio Luzzati, an emeritus investigator at the Centre for Molecular Genetics in Gif-sur-Yvette near Paris (part of the French national research agency CNRS), worked with Franklin, and remained a close friend until her untimely death from ovarian cancer in 1958. Early last year, after Klug vouched for Seckel, Luzzati provided a handwritten letter from Franklin and some old reprints of scientific articles. The idea of payment was never discussed. "I didn't charge," Luzzati says. "I don't want money."

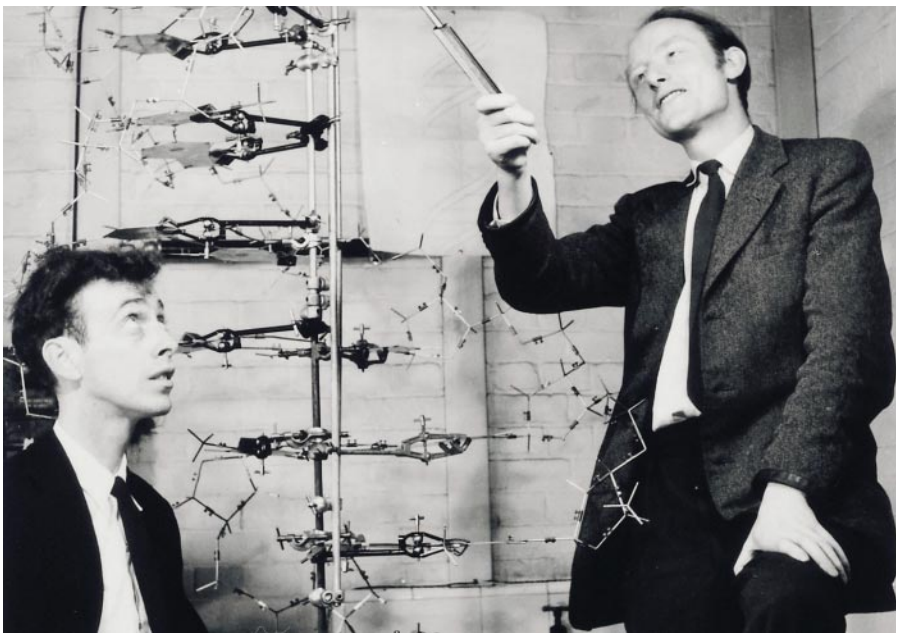
Luzzati also lent Seckel his collection of photographs of Franklin, for which there are no surviving negatives. Luzzati says Seckel was to make digitally enhanced copies for the Norman archive, and return the originals along with improved copies. But when *Nature* visited the Norman archive, Luzzati's original photographs were still in the collection. Norman says he was unaware of any agreement to return them, but he has now done so.

The big test of Norman's intentions, however, will be the ease with which interested scholars have access to the archive. The London-based journalist Brenda Maddox — wife of former *Nature* editor John Maddox — visited the Norman archive in March for several days, as part of her research for a forthcoming biography of Franklin, and was impressed by the collection. "I had several 'Eureka!' moments," she says.

Archivists who are fretting about the private ownership of such an important collection wonder how many researchers will have such moments in the future. ■

Rex Dalton is *Nature*'s West Coast US correspondent.

A. BARRINGTON BROWN/SPL



Double act: the archive contains a draft of Watson's colourful book about his work with Crick (right).

© 2001 Macmillan Magazines Ltd

Seismic sleuths

The worldwide network of seismometers is detecting some surprising events, from bouncing kangaroos to changes in climate. Larry O'Hanlon talks to the seismologists who have found unusual uses for their data.

When Russia's *Kursk* submarine went down with all hands in the Barents Sea in August last year, no one rushed to call a seismologist. But, in January this year, it was researchers from that discipline who finally ended the debate over the cause of the disaster.

Keith Koper of the University of Arizona in Tucson reported that Baltic seismic stations had recorded tell-tale vibrations from explosions aboard the *Kursk*[1]. The evidence suggested that the tragedy was caused by one of the submarine's torpedoes accidentally detonating while the craft was on the surface — an event followed by several more simultaneous detonations at depth. Russian authorities had previously claimed that a collision with an unidentified foreign submarine was to blame.

Ear to the ground: Terry Wallace's data disproved the claim that Iraq had tested a nuclear bomb.

AP; D. MCNEW/NEWSMAKERS; AP

Koper's *Kursk* paper is just one example of the unexpected uses that researchers are finding for seismological data. A global network of seismometers, established to study earthquakes and to watch out for states testing nuclear weapons, is allowing scientists to listen in on phenomena as diverse as the sonic boom created by the space shuttle as it returns to Earth, and the 'sounds' of global warming.

Detective work

Such unusual applications for seismometers came to light after what many consider to be the United States' worst home-front disaster of the Second World War. On 17 July 1944, around 5,000 tons of munitions accidentally detonated in a series of explosions at the US Navy's Port Chicago on the Sacramento River in California. The overall blast killed 320 sailors and flattened the port.

After the event, Navy officials consulted seismologist Perry Byerly of the University of California, Berkeley. The blast had registered on nearby seismometers, and Byerly was able to determine the exact timing and sequence of the different explosions[2]. Although Byerly had previously examined seismic signals from mine blasts[3], the prominence of the Port Chicago disaster meant that, for the first time, the value of such analyses attracted widespread attention.

> Seismometers are regularly used to verify claims that nuclear tests have taken place.

The drafting of the Comprehensive Test Ban Treaty (CTBT) in the 1950s propelled Byerly's work onto the global stage. The treaty, which was intended to outlaw all nuclear weapons tests, needed a mechanism by which breaches of the ban could be spotted. With this in mind, the US government in 1959 initiated the Worldwide Standardized Seismographic Network (WWSSN) as one strand of a range of strategies to spot nuclear tests. It consisted of 120 seismic stations — the first global seismic network. At the same time, the US government began pouring millions of dollars into basic seismological research, giving a tremendous boost to what was then, arguably, a peripheral field of study.

As WWSSN seismologists started to detect, measure and compile the world's earthquakes on their maps, the collected epicentres started to delineate the hitherto unknown boundaries of the Earth's crustal plates. "That really helped usher in the whole era of plate tectonics," says Terry Wallace, a seismologist at the University of Arizona and a member of the team that worked on the *Kursk* data. It was a case of sublime serendipity, with Byerly's basic research leading to a seismic network for monitoring nuclear blasts which, in turn, helped spur what proved to be the twentieth century's most significant theoretical shift in the Earth sciences.

More than 16,000 seismic stations now cover the globe. Originally, seismometers were sensitive only to vibrations at a particular frequency, so seismic stations had to use an array of different seismometers to monitor the range of vibrations that were of interest. But the development in the mid-1980s of broadband seismometers, which are sensitive to vibrations at a range of different

© 2001 Macmillan Magazines Ltd

Even the faintest vibrations have the potential to reveal seismic treasures.

frequencies, made it easier and cheaper to add new stations to the global network.

The seismometers constantly spot surprising events, from meteor impacts to rock falls or the long, slow, gravitational pull of the Moon on the Earth's crust. For the most part, these signals are a nuisance to seismologists. Noise from human activities is an increasing problem. People continue to spread over the surface of the Earth, bringing noisy machines with them. "It's what we call cultural noise," says Wallace. "Our problem is that we have an incredibly sensitive seismometer and people come and build around it."

Noise annoys

Seismologists are continually developing ways of separating the seismic wheat from the chaff. But once the interfering signals are identified and separated out, fresh uses for seismometers quickly emerge. "One man's noise is another man's signal," says Robert Uhrhammer, a seismologist at the University of California, Berkeley.

Traffic movement, for example, can sometimes be useful seismic noise. When the Yom Kippur war ended in 1973, Israel and Egypt were in dispute over tactically important passes through the Sinai mountains. Israel wanted to maintain its electronic listening station in the mountains, but the US-mediated peace agreement called for Israeli withdrawal from the region.

A compromise was reached that allowed Israel to control the station remotely if it agreed to withdraw its forces from the area. The United States installed a monitoring system, of which seismic sensors where a part, to check for Israeli troop movements near the station. Without such a system, the peace agreement may never have been signed. More recently, similar systems have been developed by the US Immigration and Naturalization Service and International Microwave, a company in East Norwalk, Connecticut, to spot illegal immigrants trying to cross the border from Mexico.

Seismometers are also still regularly used to verify, or dismiss, claims that nuclear tests have taken place. Despite decades of diplomatic wrangling, the CTBT has yet to come into force. The list of signatories has some notable exceptions, such as India and Pakistan, both of which have atomic weapons programmes. And other nuclear powers, such as the United States and China, have signed but not ratified the treaty.

But if the treaty ever does come into play, a dedicated network of monitoring stations will be ready to deploy their seismometers to help watch for nuclear tests. Many seismometers are based at university laboratories, and over half are already supplying information to the Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization, the Vienna-based body charged with setting up the system that will monitor the CTBT should it come into force.

Shortly after his work on the *Kursk*, Wallace made the news again. With his Arizona colleagues, he investigated a report in a February edition of the British newspaper *The Sunday Times* which alleged that Iraq had carried out a clandestine test of a 10-kiloton nuclear weapon in September 1989. Wallace's team managed to debunk the story by looking at the seismological record for the region. An informant had told *The Sunday Times* that there had been an underground blast beneath Lake Rezazza, 150 kilometres southwest of Baghdad. But the seismological record contained no evidence of any significant seismic activity within 50 km of the lake that day.

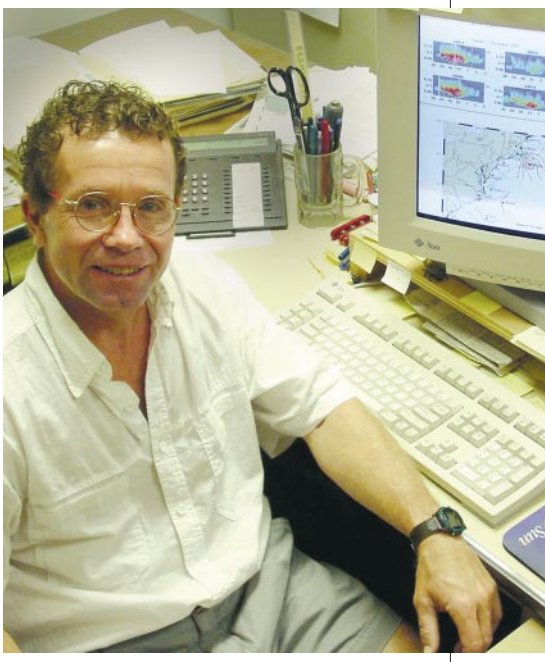
Wallace concluded that even if Iraq had, as alleged, hollowed out a huge cavern under the lake to 'decouple' the blast from the surrounding rock, the detonation of a 10-kiloton bomb could not have escaped seismic detection. Using data from decoupling experiments by the United States and the former Soviet Union as a guide, he worked out that such a blast would have caused a seismic signal with a magnitude of between 3.8 and 4.0 on the Richter scale. At the time, the seismometer network in the region was capable of detecting any signal with a magnitude of over 2.9. But only tiny seismic signals were detected on the date of the alleged test — too weak even to verify their origins.

Shakedown: as well as monitoring earthquakes, seismometers can be used to listen to waves, spot underground and undersea nuclear tests, detect the space shuttle or terrestrial traffic, view the Moon's gravitational pull on the Earth's crust, and work out the size of kangaroos.

Just in case more tangible proof was needed, there is also some compelling circumstantial evidence: the continuing existence of Lake Rezazza. According to Wallace, "The collapse of a nuclear cavity under a lake would, of course, produce a tell-tale sudden drop in the water level, if not the disappearance of the lake — a risky proposition for a country attempting a clandestine test."

Small is beautiful

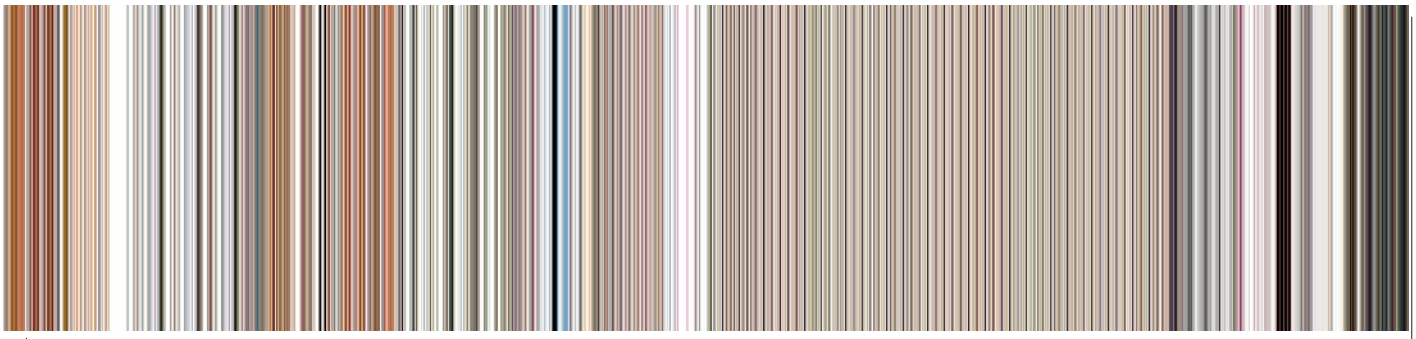
Even the faintest vibrations, known as microseisms, offer seismic treasures, says Peter Bromirski, a seismologist at the Scripps Institution of Oceanography in La Jolla, California. In the early 1990s, Bromirski was working on a postgraduate



Making a splash: Peter Bromirski hopes to assess changing wave heights using seismic data.

NASA/NEWSMAKERS; T. C. MALHOTRA/NEWSMAKERS; CORBIS

© 2001 Macmillan Magazines Ltd



JPL/NASA; AP

project for the US Navy to evaluate whether ocean-bottom seismometers could detect nuclear tests. The project is still classified, but Bromirski says that the system could have been used to monitor bomb tests in North Korea without invoking the political complications of trying to place land-based seismometers in neighbouring countries.

While processing US Navy data from seismometers placed off the coast of Oregon, Bromirski quickly realized the need to filter out ocean noises. Microseisms of ocean waves were among those noises, and he has been studying them ever since. It was a classic case of noise becoming signal, Bromirski says.

Stormy relationship

In 1999, Bromirski proposed that seismic data collected by stations along the California coastline could help fill in the gaps in climate data from the same area. Climate researchers were keen to look for long-term changes in wave heights, as many models of climate change predict that rising global temperatures will lead to more frequent storms, which would be reflected in an increase in wave heights. But ocean buoys, a key source of data on wave conditions, had only been deployed in the region since the early 1980s. By contrast, seismometers at the University of California, Berkeley, have been collecting data since about 1930. Bromirski thought that the missing wave data might be lying hidden in the seismic records.

Ocean swells approaching a coastline push down on the seabed, resulting in distinctive microseisms that travel inland. Individual waves can also be detected as they hit land. By comparing records of these microseisms from land-based seismometers with data from ocean buoys, Bromirski and his colleagues worked out how to estimate wave heights using the seismic data[4]. Applying Bromirski's technique to the seismic archives could produce seven decades of wave-height data for the northeast Pacific Ocean.

Rather than analysing the whole 70-year record, Bromirski has since used the seismic record to tease out information from earlier storms and even the first great El Niño of the twentieth century, that of 1940–41 (ref. 5). "Back then nobody really knows what the waves were like," he says. "There were no buoys and no satellites, just ship measurements and atmospheric pressure records."

Seismologists also encounter the ridiculous: kangaroos show up as they bounce along.

Ingo Grevemeyer, a seismologist at the University of Bremen in Germany, is taking a similar approach. Using a 40-year record of microseisms from a seismic station in Hamburg, Grevemeyer has shown that the northeast Atlantic Ocean has become more stormy during the past 20 years[6].

Of course, not everything forensic seismologists discover is so useful. With so much noise they also encounter the ridiculous. Kangaroos, for example, show up in the seismic record as they bounce along. "You can gauge the size of the kangaroo by the frequency of its bounces," says Uhrhammer, who has become an informal curator of seismic curiosities.

Uhrhammer has even used seismic noise to spot broken equipment. In 1999 he analysed seismic signals from a diesel generator that was powering scientific equipment on the Farallon Islands off San Francisco. Uhrhammer, an experienced mechanic, realized that one of the generator's pistons was misfiring and so was able to make repairs before the generator broke down completely.

Distance learning

Despite enjoying working with these quirks, the ultimate aim of most seismologists is still to filter them out. The best way to do that may be to escape the Earth altogether and use a network of satellite-based seismometers. Wallace proposes that a series of satellites could bounce electromagnetic signals off the Earth's surface to give real-time visualizations of, for instance, seismic energy moving along the San Andreas Fault. By moving around the Earth, the seismometers would be able to view the same point from different angles, building up a three-dimensional picture of the surface. The seismometers could watch seismic waves from an earthquake rippling over California as if the waves were from a stone dropped in a pond.

Wallace says that space-based seismometers could also monitor very slow changes in the Earth's surface that broadband seismometers miss — such as the gradual movement of the surface created by the hollowing out of underground caverns during mining activity.

Although a definite proposal has not yet been put together, and many technical challenges will have to be overcome, Wallace says that the idea has the backing of the seismological community. He believes the insights that seismologists would gain from studying quakes from such a viewpoint will lead to great advances in modelling fault zones. "I think it will revolutionize seismology," he says. "It will really allow us to see the groans of the Earth — a picture of a very dynamic planet. A picture of what's happening right now." ■

Larry O'Hanlon is a science writer in Auburn, California.

1. Koper, K. D., Wallace T. C., Taylor, S. R. & Hartse, H. E. *Eos* **82,** 37, 45–46 (2001).
2. Byerly, P. *Bull. Seismol. Soc. Am.* **36,** 331–348 (1946).
3. Byerly, P. & Wilson, J. T. *Bull. Seismol. Soc. Am.* **25,** 259–268 (1935).
4. Bromirski, P. D., Flick, R. E. & Graham, N. *J. Geophys. Res.* **104,** 20753–20766 (1999).
5. Bromirski, P. D. *Geochem. Geophys. Geosyst.* (in the press).
6. Grevemeyer, I., Herber, R. & Essen, H.-H. *Nature* **408,** 349–352 (2000).

Web links

Seismic analysis of the sinking of the *Kursk*
http://www.geo.arizona.edu/researchers/kkoper/Kursk/for_kursk.html

Detection of the space shuttle's sonic boom
http://www.galcit.caltech.edu/SonicBoom/boomslides.pdf

Monitoring of Egypt–Israel peace accord
http://www.cmc.sandia.gov/issues/papers/vannoni2/index.html

Debunking of alleged Iraqi nuclear test
http://www.vertic.org/tnv/marapr01/iraq.html

Details of the proposed Comprehensive Test Ban Treaty monitoring scheme
http://pws.ctbto.org

© 2001 Macmillan Magazines Ltd

Smart-card traffic system keeps Singapore in the fast lane

Sir — Recognizing the high economic and environmental costs of traffic gridlocks, Singapore recently implemented an electronic road pricing (ERP) system, the first of its kind in the world. Using a network of overhead gantries straddling the central business district and the approach routes and expressways leading there, the system works by remotely debiting tariffs from a smart card installed in a slot near the vehicle's windscreen. Infringement is registered by capturing the image of the vehicle's number plate. Our experience of this system, phased in during 1998 and 1999, offers some perspectives to other regions contemplating the use of this technology to manage congestion.

A major barrier to the implementation of electronic tolling is public aversion to infringement of privacy. In the passive electronic toll systems used in some parts of Australia, Canada, Europe and the United States, vehicles need to be tagged and their movements monitored for monthly billing. Singapore instead chose the cash-card system, in which direct debiting removes the need for billing and debt collection.

There needs to be political courage in implementing congestion tolls. Singapore ran a public education campaign to show that ERP is about reducing vehicle usage, not making money. Indeed, revenue has plunged by about 40% compared with the previous, manual tolling system, since fewer drivers use the tolled roads. The government absorbed the cost of fitting every vehicle in Singapore with a unit for cash cards, gave road-tax rebates and reduced customs and registration costs for new cars, so motorists saved money. This has helped public acceptance of ERP.

There is a limit to how far civil engineering can enhance road capacity in urban areas, faced with an ever-increasing number of vehicles. A congestion toll by ERP helps to price roads as scarce commodities and therefore promotes their efficient use, besides being more cost-effective than a 'manual toll plaza', or series of toll booths, where motorists have to stop and pay. The Singapore system has several advantages: there is no need for vehicles to stop or slow down as they pass the gantry; it lets large numbers of vehicles through; it allows toll charges to be incorporated based on the vehicle class; and it eliminates cash handling, thus enhancing commuter convenience and revenue security.

For expressways and approach routes leading to the city, a tiered pricing scheme operates during the peak period from 7.30 a.m. to 9.30 a.m. on weekdays. For roads in the central business district, another pricing scheme operates from 7.30 a.m. to 7 p.m. The system has allowed various pricing schemes to be tested, so that optimal speed levels are achieved on expressways (45–65 km per hour) and roads in the central business district (20–30 km per hour) by reducing the number of vehicles using tolled roads during peak hours.

Unlike the previous system, under which daily or monthly permits allowed an unlimited number of journeys, the present pay-per-trip mechanism discourages extra car use. The system motivates drivers to try quieter roads, use public transport or car-sharing schemes, and drive in off-peak periods where possible. The ERP experiment is currently being monitored before extending it to more freeways in Singapore.

Leo Tan Wee Hin, R. Subramaniam
Singapore National Academy of Science, 15 Science Centre Road, Singapore 609081

Mathematical model could clarify arms race

Sir — The US government's recent decision to enlarge its missile-defence systems, and the response of other countries in Asia which have perceived this as a threatening move, raise the question of whether a new global arms race is beginning. Mathematical models about the relationship between armaments and conflict need to be discussed openly by scientists, ideally in a structured context similar to the way in which climate-change models are discussed, to help governments make the most appropriate strategic decisions.

Lewis Fry Richardson developed a simple model of coupled differential equations to show how expenditure on armaments by antagonistic nations had grown exponentially before the First and Second World Wars. This led him to predict — in a letter beginning "As *Nature* has encouraged scientific workers to think about public affairs …" (*Nature* **135,** 830; 1935) — that there would be a second world war when the armaments of both sides became dangerously numerous. War could be avoided, he warned, only if the leaders of the arming countries took immediate and decisive action.

By 1951, Richardson had developed and refined his model, and *Nature* published his letter (*Nature* **168,** 567–568; 1951) in which he asked whether the third major arms race of the twentieth century would also inevitably lead to a world war. In some respects his refined model predicted the ending of the US–Soviet cold war, though it has been argued that his equations do not strictly allow such an interpretation (see I. Sutherland, *Collected Papers of Lewis Fry Richardson* Vol. 2, Cambridge University Press, 1993).

Most countries reduced their stock of armaments during the 'new world order' period between 1989 and 1999, with the exception of China and India. The question now is what kind of prediction would result from conflict/armament models such as Richardson's when applied to the present situation. As things stand, the model equations tend to show great sensitivity to small variations in relative armaments at this critical stage, so much work is needed. It is to be hoped that concerned "scientific workers" and others are convinced that the present international situation is potentially serious, and that every relevant method should be applied to the prediction of events and the exploration of constructive strategies.

Political analyses are necessary, but are not a complete substitute for scientifically based quantitative and probabilistic evaluations of complex outcomes and possible remedial actions.

Nowadays, when governments are more open about their use of mathematical models for systems such as the economy, climate change and spread of disease, it could be argued that they should also be open about how they are applying mathematical models to the international strategic situation. Those models would be more effective in influencing events if they were widely applied and accepted, especially in countries where crucial political and defence decisions are now being considered.

J. C. R. Hunt
Department of Space & Climate Physics, University College London, Gower Street, London WC1E 6BT, UK

More light shed on the Mauna Kea controversy

Sir — Your News report "Astronomers bargain for use of 'sacred' site" (*Nature* **410,** 1015; 2001) was an excellent overview of the cultural dimensions of the controversy on Mauna Kea, Hawaii. However, your story did not mention two other key issues fuelling local animosity towards the observatories — past mismanagement and proposals to limit public access.

In 1998, a legislative audit concluded that, during its 30-year tenure on Mauna Kea, the University of Hawaii's Institute for Astronomy (together with the state's land

© 2001 Macmillan Magazines Ltd

board) had "failed to develop and implement adequate controls to balance the environmental concerns with astronomy development". The audit also confirmed islanders' charges of broken promises, violated leases and ignoring of environmental laws.

After the audit, the university's president and the Institute for Astronomy blamed the public for problems on the mountain top, and proposed access restrictions that were later included in the university's second, flawed master plan, discussed in your report. All this has made the islanders even more upset.

Given this history, the quote by NASA's Keck programme manager, John Lee, that the organization does not want to "poison the well" for astronomy on Mauna Kea, seems ironic to many islanders. NASA and the international astronomy community need to understand that those 'toxins' were introduced by the Institute for Astronomy and the observatories, and will have to be removed by them before anything else is built on the mountain.

Nelson Ho
Hawaii Chapter of Sierra Club, 32 Kahoa St Hilo, Hawaii 96720-2206, USA

A new approach to global book distribution

Sir — As pointed out by Keese in Correspondence (*Nature* **410,** 1021; 2001) and in your current web debate (http://www.nature.com/nature/debates/e-access/index.html), lack of access to scientific information is a widespread problem. I have experimented with a means for obtaining free books and distributing them to individuals and libraries, and believe this technique could be widely used.

The cost of extending print runs to produce extra copies of books is reasonably small — comparable to typical royalties. When I completed a conservation textbook for Blackwell Scientific, the publisher agreed to my suggestion that, in lieu of royalties, it would provide me with a free copy for each one sold. My aim was to send the book out to practising conservationists who would otherwise be unlikely to obtain it. The online bookshop nhbs.com (http://www.nhbs.com) offered to coordinate the distribution, and the Christensen Fund kindly paid for the postage.

As this scheme became known, we received many requests and suggestions for recipients (we welcome further suggestions — please contact me at w.sutherland@uea.ac.uk). Within a few months of publication we were able to send out more than 1,250 copies of the book to people in 126 developing countries.

Such schemes could be widely applied to scientific and technical books, and everyone will benefit if the following three suggestions are followed. First, authors could ask publishers to provide all or some of their royalties in the form of extra copies. Authors write to be read, and this procedure greatly increases the distribution and hence the readership of a book. Second, academic societies and grant-awarding bodies could allocate funds for posting books, a very cost-effective means of promoting a subject. The distribution could be carried out at cost price, organized either through an academic body, through the author's institution or through a commercial organization. The distributor can gain some excellent publicity by doing this. Third, publishers could advertise the fact that they are happy to proceed with such arrangements.

The charity Book Aid International (www.bookaid.org; info@bookaid.org) has also expressed an interest in distributing suitable books through its network.

The scheme offers clear advantages to publishers, as it improves their image and may attract authors. In the current case, the publisher believes the offer has increased sales.

William J. Sutherland
School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Improving taxonomy for us and the other fishes

Sir — Lewin in his engaging Concepts essay (*Nature* **410,** 637; 2001) asserts that unless "demanded by expediency", recent changes in taxonomy "muddle our language, and could introduce confusion in our libraries". Two of his examples involve traditional groups that are unambiguously paraphyletic (each of them includes an ancestor but not all of its descendants), and they nicely illustrate why monophyletic taxa (those comprised of a common ancestor and all of its descendants) are preferable for teaching, research and influencing public attitudes towards nature.

Lewin claims that molecular phylogenetics has led to "a host of new problems": for example, "Because crocodiles are believed to be more closely related to birds than they are to turtles … are sparrows just feathered reptiles, and does ornithology become merely a branch of herpetology?". That birds and crocodiles are each other's closest living relatives was a consensus view long before the advent of DNA sequencing, as has the conclusion that reptiles (including birds) are more closely related to mammals than they are to amphibians. Herpetology, defined monophyletically as the biology of tetrapods, thus does incorporate mammalogy and ornithology, but this no more makes sparrows "just" feathered reptiles than it reduces bats to "just" winged mammals; ornithology is no more "merely" a branch of herpetology than the latter is "merely" a part of vertebrate biology.

More important, traditional failure to taxonomically formalize a close relationship between crocodilians and birds obscures their similarities in parental care, acoustic communication and various internal traits. Perhaps that failure also explains why some recent analyses of life-history evolution in each group fail to cite work on the other, and why they also ignore a substantial literature on reproduction in closely related, extinct archosaurs.

Lewin correctly notes that the coelacanth *Latimeria* "is closer to humans than it is to herrings", but then complains, "so what price ichthyology?". Actually, the incredible panoply of fishes is done an injustice by traditional recognition of two basal taxa, one of them spectacularly successful (there are more than 20,000 species of ray-fins) and the other usually viewed as an evolutionary dead-end (*Latimeria* and lung-fishes).

A monophyletic perspective enriches ichthyology by, instead, emphasizing that the adaptive radiation of fishes includes two more-or-less equally speciose subclades, one largely confined to water (the ray-fins) and the other represented by both aquatic and terrestrial forms (coelacanths, lung-fishes, tetrapods).

Of course, one could argue that tetrapods are special because they move about on land and have changed a lot, but then so have walking catfish, swamp eels and mudskippers, all of which we appropriately leave in with other ray-finned fishes.

Lewin disapproves of new names and altered meanings for old names, whereas my experience is that students are inspired by the underlying rationale for such changes. Lay people likewise are intrigued by behavioural similarities between crocodilians and birds, and amused by the realization that we are all fishes. Rather than hindering biology, increasingly accurate and phylogenetically based taxonomy promotes the study and appreciation of life's diversity.

Harry W. Greene
Department of Ecology and Evolutionary Biology, Cornell University, Ithaca, New York 14853-2701, USA

© 2001 Macmillan Magazines Ltd

The philosopher's child

The death of Darwin's daughter had profound effects on his work.

Annie's Box: Charles Darwin, His Daughter and Human Evolution
by Randal Keynes
Fourth Estate: 2001. 346 pp. £16.99

Bruce Weber

The approach of the millennium brought with it a rash of speculation as to who was its most influential individual. Among scientists, strong cases were made for Isaac Newton and Charles Darwin. Go into any large bookshop today and you will find a range of books on Darwin, darwinism and evolution or with Darwin somehow worked into the title. And in America there are also books on the evolution–creation debate. As Randal Keynes points out, Darwin's influence is still strongly felt in a way that the legacies of those other icons of nineteenth-century culture, Marx and Freud, are not. Indeed, interest in Darwin, as a person as well as a thinker, is at an all-time high.

Until recently, Darwin was generally viewed as a bit of a plodder, not particularly bright and with no interest in abstract or philosophical thought. But the surge of Darwin scholarship that followed the centennial of the publication of *On the Origin of Species* began to shift perceptions. Re-evaluation of the corpus of Darwin's writing has revealed a man of keen and profound intellect who was sophisticated in the epistemology and metaphysics of science as well as in the philosophy of mind.

The publication of Darwin's notebooks and correspondence has given us insight into his thought and spawned the 'Darwin Industry'. For example, we now understand how much Darwin was influenced by John Herschel's *A Preliminary Discourse on the Study of Natural Philosophy*, which he read when it was published in 1830. Darwin wanted to bring the study of natural history within the orbit of natural philosophy. He tried to do this, not by reducing biology to newtonian physics, but rather by identifying the natural laws governing biological systems. And for this, he used a methodology that was consistent with Herschel's characterization of newtonian science.

Two major Darwin biographies have been published in the past decade — *Darwin: The Life of a Tormented Evolutionist* (W. W. Norton, 1994) by Adrian Desmond and James Moore, and *Charles Darwin: Voyaging* (Princeton University Press, 1996) by Janet Browne. These books summarize much of what we have learned about Darwin, placing him in his time and social milieu. *Annie's Box* is an important complement to these biographies and provides us with a more intimate and personal portrait of the man. It focuses on Darwin's relationship with his eldest daughter Annie and how her death at the age of ten affected his subsequent research.

Long past: Annie, and her writing case, which still survives.

Keynes, an Oxford-educated anthropologist and retired civil servant, draws on his background, and on material held by the Darwin family (he is the great-great grandson of Charles Darwin), to explore the origin and character of Darwin's intellectual preoccupations and creativity. Although not a trained historian, Keynes demonstrates a rigorous scholarship based on his own substantial research and ably distils that of others.

Perhaps because he is not an academic historian, Keynes writes in a lively style that will engage the general reader. By using apt quotations from published and unpublished Darwin material, and from popular literature and sermons of Darwin's time, Keynes weaves a rich tapestry that gives the reader a sense of the attitudes and assumptions of the Darwin family and their class. Thirty-four illustrations, 11 of which have never before been published, give visual support to Keynes's claim that Darwin's observations of apes, savages and babies were pivotal to the development of his ideas.

Even as Darwin was speculating on

© 2001 Macmillan Magazines Ltd

natural selection in 1838 in his notebooks, he was making a career in the scientific circles of London on the basis of the vast array of material from the *Beagle* voyage. He was elected secretary of the Geological Society, and at a meeting of the society in February 1838, the president, William Whewell, remarked that the many fossils Darwin had observed in South America led to "profound and enlarged speculations" on "the diffusion, preservation, and extinction of animal races".

Keynes argues that, under Herschel's influence, Darwin was indeed engaged privately in "profound and enlarged speculations" on universal laws of biology that might account not only for extinctions but also for adaptations, speciation and common descent, and that he took Whewell's words as provocation and challenge. He also sought to integrate his experience of the 'savage' men of Tierra del Fuego and to contemplate the possibility that humans had evolved.

As he observed the behaviour of Jenny, the orang-utan on display at the Zoological Society in London, he began to speculate on how human behaviour and mind might have evolved from apes. He experimented with Jenny's response to music and to seeing herself in a mirror. These were experiments he later performed on his own small children, including Annie, as he carefully observed their behavioural development. Keynes presents Darwin's observations of the 'natural history of babies', not as the detached studies of an increasingly doubting sceptic, but rather as the emotionally engaged experiences of a loving father who sought to see life's development as a whole.

Of all his children, Annie was the one to whom Darwin felt closest. As he dissected barnacles in the late 1840s, she would often stand beside him and then join him on his daily walks. Her illness and death in 1851 were devastating for Darwin, especially as he feared that her disease might have been inherited, for her symptoms of severe digestive distress reminded him of his own chronic problem. But, as Keynes points out, it is most likely that Annie had tuberculous peritonitis, common at that time, and unrelated to her father's digestive distress.

While Emma Darwin saved Annie's writing case (the box of the title), Charles wrote a memorial to Annie and then sought solace in grappling with the origin of species. He found no comfort in Christianity and stopped attending church services with the family. God had begun to recede from his life. And this made it easier for him to face the cruelty and waste that seemed inherent in the 'laws of life' as he worked on the *Origin of Species*.

Keynes finds passages in the chapter entitled "The struggle for existence" that echo the language in Darwin's memorial to Annie. He believes they reflect Darwin's changed perception of the "face of Nature", especially when the earlier, emotionally neutral 'wedge' simile for natural selection described in his 1838 notebook is compared with the more violent one in 1859. The image of wedges forcing out other wedges to create gaps in the economy of nature was transformed into the "face of Nature" as "a yielding surface, with ten thousand sharp wedges packed close together and driven inwards by incessant blows, sometimes one wedge being struck, and then another with greater force".

Although Darwin made only passing reference to human evolution in the *Origin*, it was in his conceptual scheme from 1838. But the reaction of even many of his supporters to the modest comments about human evolution was disappointing. Charles Lyell could not accept human evolution and Alfred Russel Wallace ultimately drew the line at the evolution of the human mind. In *The Descent of Man* (1871) Darwin took up the question directly, further developing his themes in *The Expression of the Emotions in Man and Animals* (1872) and *A Biographical Sketch of an Infant* (1877). The issues raised in these works have retained their vitality and still provoke controversy. Keynes argues reasonably that the importance Darwin placed on memory in human nature and on morality in these works flowed from his memories of Annie.

Keynes sees in Darwin's relationship with his daughter threads that tie together disparate aspects of his life and thought. He makes a compelling case and reveals an aspect of Darwin that should be incorporated into our total picture of him. The writing is graceful, the illustrations are apt and affecting, and the thesis is convincingly presented. ■

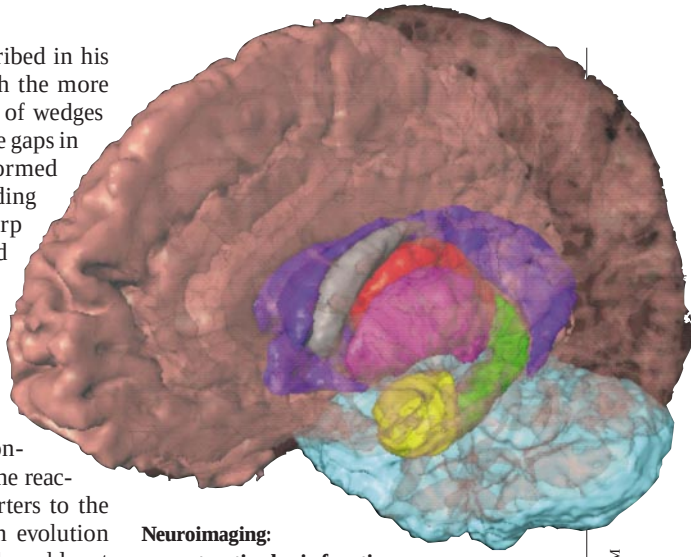
Bruce H. Weber is in the Department of Chemistry and Biochemistry, California State University, Fullerton, Fullerton, California 92834, and in the Division of Science, Bennington College, Bennington, Vermont 05201, USA.

Scanning the mental continuum

Brave New Brain: Conquering Mental Illness in the Era of the Genome
by Nancy C. Andreasen
Oxford University Press: 2001. 390 pp. $29.95

Robert Plomin

Psychiatry has come a long way since 1984 and *The Broken Brain* (Harper & Row), Andreasen's previous book for a general readership. That book described how psychiatry had been shaped by three models: psychodynamic, behavioural and biological. The psychodynamic, or freudian, perspective was then still dominant, but the book predicted a shift towards a biological model. *Brave New Brain* heralds the arrival of a biological model that combines neuroscience and genetics.



Neuroimaging: reconstructing brain function.

IOWA BRAINS SOFTWARE TEAM

The first half of the book provides excellent introductions to neuroscience, brain imaging and genetics. The second consists of outstanding overviews of what is known about the neuroscience and genetics of the major psychiatric disorders of schizophrenia, mood disorders, dementia and anxiety disorders. The descriptions of the history and neuroscience of medication are especially good. Although the book is intended for a general readership, it will be valuable for students and for scientists who are not expert in all three areas of neuroscience, genetics and psychiatry.

You feel you are in safe hands — and you are. Andreasen is editor-in-chief of *The American Journal of Psychiatry* and winner of the prestigious US National Medal of Science. She is also a practising psychiatrist and a neuroscientist best known for her research on neuroimaging.

Those who worry about biological determinism in a book on neuroscience and genetics can be reassured: Andreasen is not a biological determinist. She emphasizes the important role of non-genetic factors in mental illness and the plasticity of the brain: "Our brains are constantly rewiring themselves so that we very literally 'change our minds'." Brains can be changed not just by pharmacology but also by psychotherapy.

Andreasen mounts her soapbox to argue that the main villain in the dehumanization of psychiatry is not biological determinism but the economic revolution in the provision of health care in the United States over the past two decades and now increasingly in the United Kingdom. "Its impact on psychiatry has been especially bad since the economic revolution follows principles that are unabashedly nonhumanistic. Spending time talking to patients, or listening to them, is now considered an expensive luxury to be

© 2001 Macmillan Magazines Ltd

avoided whenever possible. But talking and especially listening are central to good psychiatric evaluation, and they form the basis for most psychotherapy." You will know what she means if someone you love suffers a mental illness.

Psychiatry is where neuroscience is making the strongest links with the genetics of variation, that is, the differences in DNA between individuals of the same species. Neuroscience has traditionally focused on species-typical, or normative, brain phenomena. Psychiatric genetics, on the other hand, focuses on variations in these species-typical themes as psychiatrists try to understand why some people become mentally ill whereas others do not. During the past few decades it has become clear that hereditary differences between people make an important contribution to differences in susceptibility to mental illness.

Neuroscience aims to understand how the brain works generally — for example which bits of the brain light up under neuroimaging when particular tasks are performed. Until now, genetics has been involved in neuroscience only in terms of rare single-gene mutations in humans or gene targeting in mice, in which mutations are induced as a way of identifying normal brain processes. This approach tends to treat all members of a species as if they were genetically identical except for a few rogue mutations that screw things up. So from the species-typical perspective of neuroscience, mental illness is a broken brain.

There has, however, been a quiet revolution in the genetics of normal variation among individuals. No longer are researchers looking for *the* gene for schizophrenia or for any other common disorder, whether behavioural or not. It is now generally accepted that common disorders are influenced by multiple genes. Such genes are often called quantitative trait loci (QTLs), because the implication is that if multiple genes are involved, the trait will be distributed quantitatively as a continuum, like height. From this perspective, common disorders may be merely the extremes of the same genetic and environmental factors that create variation throughout the normal distribution. In other words, there may be no disorders as such, just the extremes of quantitative dimensions. When this QTL perspective reaches neuroscience, it should foster research on normal variations in brain processes. Then mental illness will no longer be thought of as a broken brain. Just as there is no single human genome, there is no single human brain.

Although Andreasen mentions this new view, she has not fully come to terms with it. This is understandable, as psychiatry itself is just beginning to tackle its implications — neuroscience hasn't even begun to think about it yet. More generally, genetics and neuroscience are not yet integrated. This can be seen in Andreasen's book despite her valiant effort to bring them together. I hope that in her next book she will be able to say that her prediction about the integration of neuroscience and genetics in psychiatry has become reality. ■

Robert Plomin is at the Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK.

In search of perfection

Quantum Mechanics: Symbolism of Atomic Measurements
by Julian Schwinger (edited by Berthold-Georg Englert)
Springer: 2001. 484 pp. £34, $49.95, DM98

Alastair I. M. Rae

Julian Schwinger shared the 1965 Nobel prize for physics with Richard Feynman and Sin-Itiro Tomonaga for their invention of the theory of quantum electrodynamics. This theory extends the quantum ideas that underpin our understanding of how sub-atomic particles behave to the electromagnetic field itself.

Feynman is probably the best known of the three Nobel prizewinners for two main reasons. First, he relied heavily on his profound intuition for physics, and the concept of the 'Feynman diagrams', which are widely used in theoretical physics today, was invented in this context. Second, Feynman is well known among physicists for his 'red books' — a series of texts published in the 1960s which cover much of the basic physics taught to students. In particular, the third of these covers quantum mechanics and was based on notes for a series of lectures delivered at the California Institute of Technology. This is now a classic text and many of today's physicists learned their quantum theory from this book.

Schwinger had also planned to publish a text on quantum mechanics, having taught it in the 1950s at Harvard University. But he held back because he "had not yet found the perfect way to teach quantum mechanics". He died in 1994. Nevertheless, notes to his lectures were published in 1955 and again in 1990, and they have now been developed into a full textbook by Berthold-Georg Englert.

The contrast between this book and Feynman's reflects the different ways in which they contributed to their Nobel-prizewinning research. Open Feynman at random and you will find a page of text containing half a dozen equations; a typical page of Schwinger's book is full of mathematics with a few lines of text.

An exception is the opening chapter, which sets out what Schwinger believes are the "deep philosophical lessons to be learned … about the foundations of the subject". The chapter starts with an analysis of classical physics in its particle newtonian formulation and Maxwell's theory of the electromagnetic field. These no longer apply in the atomic regime, where it is found that classical electromagnetic fields (such as light) have particle properties (photons) whereas electrons, traditionally thought of

AIP EMILIO SEGRÈ VISUAL ARCHIVES

Julian Schwinger: philosophical lessons from quantum mechanics.

© 2001 Macmillan Magazines Ltd

as particles, have wave properties. Quantum theory confines itself to predicting the probabilities of different outcomes of experimental measurements (usually of particle properties), which can be calculated using the wavefunction.

Schwinger's philosophical approach would be considered pretty conventional today. He describes and approves of Bohr's principle of complementarity, which claims that certain observations can never be made simultaneously — for example, an electron can never be seen as a particle and a wave at the same time. And, in common with many of his contemporaries, he shows little awareness of the fundamental 'measurement problem' in quantum mechanics, which Schrödinger had exposed with the help of his famous cat and which many believe remains unresolved to this day.

To determine a set of mathematical tools for describing quantum mechanics, Schwinger analysed many variants of the Stern–Gerlach experiment. This experiment, which involves shooting a beam of atoms through a magnetic field, measures the quantized angular momentum of an atom. Schwinger's results enabled him to establish the now-familiar algebra of operators and state vectors. This approach contrasts with the more traditional one, which starts with the Schrödinger equation and the wavefunction for a free particle. Yet in many ways it parallels the treatment in Feynman's red book, although it is far more mathematical and formal. It is therefore surprising to come across an explicit derivation of an elementary trigonometric identity. Did Schwinger anticipate the lack of fluency in basic mathematics that is so common among modern students?

The scope of the later sections of the book is extensive, though not exhaustive. The section entitled 'Quantum dynamics' treats the standard problems of the harmonic oscillator in one, two and three dimensions. And the case where the oscillator is driven by a time-varying force is also included. The hydrogen atom is discussed, using an interesting analogy with the two-dimensional oscillator problem. In a section of the book that covers 'Interacting particles', we find an exhaustive discussion of the two-particle Coulomb problem and a chapter on identical particles that introduces quantum field theory. The latter topic is developed further in a final chapter on electromagnetic radiation.

So, do we have the "perfect way to teach quantum mechanics"? Probably not, if only because perfection is generally unattainable. What we do have is an interesting example of the way the subject was understood by one of the greatest physicists of the last century. ■

Alastair I. M. Rae is in the School of Physics and Astronomy, University of Birmingham, Birmingham B15 2TT, UK.

Science in culture

Mellifluous mathematics

Helaman Ferguson's topological sculpture.

Martin Kemp

There are artists who engage effectively with science. And there are scientists involved at a high level in various arts. But there are few who can proffer such a unified *curriculum vitae* in art and science as the mathematical sculptor and sculptural mathematician, Helaman Ferguson.

Originally apprenticed to his stonemason adopted father, Ferguson studied painting, sculpture and mathematics at Hamilton College in New York , and subsequently gained a doctorate in mathematics from the University of Washington in Seattle. He taught mathematics for 17 years at Brigham Young University in Utah and was co-discoverer in the late 1970s of the first generalized euclidean algorithm. His sculptures are founded on advanced topological theorems of plastic geometry, working variations on forms with such picturesque names as Alexander's horned wild spheres, toruses with cross-caps and Thurston's hyperbolic knotted wye.

Such a literally calculated approach might lead us to expect that Ferguson's sculptures would be cool abstractions in hard stone or intellectual conundrums in gleaming metal. Yet, in reality, they are beguilingly sensual and organic, subtly responsive to light and the mobile viewer, stimulating our tactile instincts in a remarkably seductive manner.

A characteristic example, on show at the National Science Foundation in Arlington, is *Whaledream II*, carved from a wonderful piece of Carrara marble. It assumes the form of a horned wild sphere in which the branched outer contour of the sphere's surface actively encompasses intricate spaces (in contrast to a standard sphere in which the inside of the surface simply surrounds a spherical space in the normal way). When we look at the complex mathematical form arising from this intellectual bedrock, we are presented not with a dry demonstration but with a mellifluous series of embracing curves that speak serenely of physical and emotional harmony. Cool calculation generates warm feelings. How can we explain the apparent paradox?

At one level, the reconciliation is affected by Ferguson's sensual intuition as a carver and skilled modeller, a master craftsman who delights in the inherent responsiveness of different materials, alternately resistant to stretching and compression. At the morphological level, however, there is no paradox at all. Ferguson has shown how topology stands alongside other geometries that generate forms analogous to those seen in the natural world. Alongside D'Arcy Thompson's more euclidean formulae of growth and form, and, more recently, alongside the new mathematics of complexity, chaos theory, fractals and self-organized criticality, the repertoire of shapes is aligned generically with such natural structures as shells, horns, tendrils and soft-bodied marine organisms. He is inventing a form of living engineering that exists in parallel to the organic architecture of nature.

A horned wild sphere: Ferguson's *Whaledream II*.

Ferguson recalls a recent demonstration of the probity of his structural principles. He points to a photograph in the National Science Foundation exhibition. It depicts "a negative gaussian curvature carving I did in 20 tons of snow at the ski resort in Breckenridge, Colorado [for an international snow sculpture competition]. I carved the snow walls down to about 4 inches thick. The next week Breckenridge had a heat wave and all the other sculptors' works sagged and imploded. Mine just got thinner and more gracile, retaining its form. The reason seems to be that on a negative gaussian curvature carving every point is a saddle point, every point is a keystone of a fabric of arches. The snow event was a proof of a simple but beautiful mathematical principle."

It is clear that we react intuitively at a very deep level to the morphological 'rightness' that is eloquently embodied in forms that are at once invitingly complex and elegant. Ferguson's creations work compellingly with our fundamental human instinct for the special kinds of mathematical order that are embedded in the world we see around us. ■

Martin Kemp is in the Department of the History of Art, University of Oxford, 59 George Street, Oxford OX1 2BE, UK.

Helaman Ferguson's sculptures and photographs are on view in "The Statue in the Stone" exhibition at the National Science Foundation, Arlington, Virginia, until 31 July 2001.

© 2001 Macmillan Magazines Ltd

Bad words

Inappropriate terms can confuse rather than aid understanding.

Melvin Konner

One of the problems with behavioural science is that for generations its glossaries have been leaky, and some of its worst coinages have seeped into general circulation. When these reach the ears of 'real' scientists, the sniggers can be heard across the quadrangle. Of course, physicists have had their aether and chemists their phlogiston. But solecisms from behavioural science somehow seem funnier, and harder to get rid of.

Take *anal fixation*, for instance. Sigmund Freud invented it in the early twentieth century, and 'anal' still commonly means too orderly, too conscientious, or tending to collect and hoard. Freud's claim was that all children pass through an anal stage of development — roughly, toddlerhood — when faeces are fascinating and their control a goal. By curious dynamics, these concerns morph into mature adult motives — normal cleanliness or organization. But if your child becomes 'fixated' at the anal stage, watch out. You could end up with an accountant or a hygienist.

Not to worry, though. There is little evidence that a process akin to fixation ever occurs in childhood, least of all in the anal realm. And if we sometimes link coins and *ca-ca* in our minds, it is probably because Freud's phrase has long since become folklore. Fortunately for civilization, orderly people do exist. The trait runs in families, so it is probably partly genetic, although it undoubtedly also rests on patterns of child-rearing. But potties and dirty nappies? Hardly.

If Freud invented a mental process out of whole cloth, the US psychologist B. F. Skinner went to the other extreme with his non-view of the brain. He and his followers grandly ignored brain science, which had become pretty heavy sledding. But some kind of justification for ignoring it was needed, because the brain is, after all, the organ that generates behaviour. So the brain became a *black box*, and for all intents and purposes an empty one. It didn't faze the behaviourist, because that paragon of rigour measured input and output, and discerned laws relating the two. As the laws of learning were considered to be universal, the stuff inside the skull was paradoxically best understood by ignoring it.

Alas for simplicity, the laws were not universal. Even rats learned some things better than others, linking odours to nausea much more easily than to shock, for example. Species also did a lot of phylogenetic backsliding in the elegant learning paradigms, misbehaving badly according to their instinctual lights — animals can be trained to exchange coins for food, but pigs will push them around with their snouts, raccoons rub them together and chickens peck at them. This 'instinctive drift' distracts them from the task they have learned. Worse, reward itself turned out to be a brain function, drawing erstwhile behaviourists deep into the black box to find out what this pleasure thing was. Now geneticists with their knockouts are picking learning apart at the macromolecular level — and, sure enough, genes for learning do exist.

Meanwhile, social-cognition theorists have come up with a phrase inferential enough to make one almost long for the black-boxers: *theory of mind*. Freud sought one, Skinner assiduously didn't, and most people don't bother to ask themselves whether they have or need one. Yet there is serious debate as to whether chimpanzees or four-year-olds have a theory of mind. Closely inspected, the phrase seems to mean something like perspective-taking or, when mutual, intersubjectivity. True, a four-year-old can see and act on another person's perspective whereas most three-year-olds can't.

This is fascinating stuff and something we need to understand. But a term such as 'theory of mind' simply stands in the way. It makes for catchy article titles but conveys no meaning. Is the maturing orbitofrontal cortex newly able to calm an impulsive and self-centred limbic circuit? Is there a down-regulation of some neurotransmitter receptor, allowing a younger form of social mirror-imaging to grow into identification and parallel perspectives? As long as we are playing with pretty word-coins that substitute for brain functions, we will never know.

Social scientists, too, have their bad words, the worst probably being the *organic* view of society. Here, instead of a turning away from biology, we have an attempt to embrace it. But this inappropriate metaphor ends up keeping natural science at bay. Society is no organism, people are not cells. Cells, unless they are sperm, eggs or cancerous, cannot even partly secede from the organism, whereas individuals are constantly reappraising their membership in society. Many secede, and those who stay cooperate very imperfectly, because cooperation is not their main goal. What are called social pathologies are not derangements of an ideal, harmonious unit, but by-products of the normal flux of a gaggle of individuals going their own way. This is so because evolution by natural selection decrees it, and most social scientists have not come to terms with the fact that people are evolving animals.

This is not to say that biologists have all the answers. Understanding must come at every level of integration, each with its own laws. Physics cannot explain biology, and complexity theory tells us that we will not get the answers we need about mind and culture merely by reasoning upward from below. Still, you can't have a science that doesn't make some sort of integrative peace with the neighbouring, more fundamental sciences in the loose but still meaningful hierarchy of nature. We social scientists should stop our silly word-play and hit the biology books. ■

Melvin Konner is in the Department of Anthropology, Emory University, Atlanta, Georgia 30322, USA. He is the author of the completely revised edition of The Tangled Wing: Biological Constraints on the Human Spirit, *to be published in September by W. H. Freeman.*

Burrhus Skinner: considered the brain as an effectively empty 'black box'.

BETTMANN/CORBIS

> We social scientists should stop our silly word-play and hit the biology books.

© 2001 Macmillan Magazines Ltd

A watery arms race

Victor Smetacek

Imagine yourself in a light forest looking upwards, seeing in your mind's eye only the chlorophyll-bearing cells of the canopy floating in mid-air, free from the attachment of leaves, twigs, branches and trunks. Now forget the forest and the trees, and see only blurred clouds of tiny green cells obscuring the blue sky beyond. You are looking at a phytoplankton bloom of a density typical of lakes and coastal oceans. Forests and algal blooms fix about the same amount of carbon — a few grams per square metre per day — because both are based on essentially the same photosynthetic machinery, fuelled by chlorophyll *a* in chloroplasts, the descendants of free-living cyanobacteria that have since evolved into plant organelles by endosymbiosis.

Chloroplasts provide their host cells with food in return for resources and protection. The land was colonized by one type of chloroplast/host cell, and the evolution of its various life-supporting systems is, from a human perspective, a straightforward success story: from algal slime to tropical rainforest. Indeed, the sole function of land plants, as considered in the thought experiment above, is to provide the chloroplasts with water and nutrients and give them access to light.

MICROFIELD SCIENTIFIC/SPL

Safe: diversity in plankton has its roots in defence.

Competition for resources and resource space has shaped the evolution of form and function in terrestrial vegetation. Can one apply the same evolutionary criteria to the other main plant life-form on our planet — the free-floating plankton of the pelagic realm? The phytoplankton bloom is suspended in a soup of resources, circulated by the wind within the sunlit surface layer. Its chloroplasts are provisioned by this viscous medium and do not require life-supporting hosts. Moreover, a striking feature of pelagic systems is the recurrent pattern of annual species succession. This is different from succession in land plants because the various stages, dominated by characteristic phytoplankton species, last for only a few weeks. There may be competition between species at the same stage for light and nutrients, but hardly at all between species of different stages. Apparently, space-holding plankton has not evolved.

So what other forces shape plankton cells, and are they the same as those that drive succession? Photosynthesis in plankton is spread across about ten different divisions, as separate from one another as land plants are from animals. Many of the lineages have species that function as algae ('plants') or as ingestors of particles ('animals'); many species do both. Generally, species with chloroplasts look no different to their relatives without them — cell shape does not reflect the mode of nutrition. Properties of the host cell, including shape, must do more than improve the photosynthetic efficiency of chloroplasts. Indeed, the enormous diversity of lineages and shapes present in unicellular plankton has defied explanation.

Although adoption by a host must have imposed many changes on the chloroplast, one main function of host cells is to protect chloroplasts against attack. The many mechanical and chemical defence systems evolved by land plants have elicited an equally heterogeneous arsenal of attack systems among their enemies, ranging from viruses to fungi, insects to elephants. Defence systems need to be deployed at the level of the leaf and are therefore not reflected in gross morphology, but they can be expensive. Hence there are fast-growing and slow-growing plants, all fuelled by chloroplasts, but differing in the degree of investment in defence.

The range of defence systems in plankton is only now coming to light. The size range of phytoplankton spans three orders of magnitude, but that of its predators spans five orders, from micron-scale flagellates to shrimp-sized krill. Pathogens (viruses and bacteria) pose a further challenge. Most predators and pathogens feed or infect selectively. Smaller predators hunt individual cells, whereas larger ones use feeding currents, mucous nets or elaborate filters to collect them *en masse*. Captured cells are pierced, ingested, engulfed or crushed, but have evolved specific defence measures. They can escape by swimming or by mechanical protection; mineral or tough organic cell walls ward off piercers or crushers. In adapting to deterring predators, cells have increased in size, formed large chains and colonies, or grown spines. Noxious chemicals also provide defence.

Plankton

"Planktonic evolution is ruled by protection and not competition. The many shapes of plankton reflect defence responses to specific attack systems."

Obviously, none of these defence mechanisms conferred by host cells provides universal protection to chloroplasts. Most phytoplankton cells are eventually eaten or succumb to pathogens. Rapid fluctuation in population size favours survival fitness, more cycles and hence more adaptation to attack. If the carbon fixed by planktonic chloroplasts is invested mainly in this biological 'arms race', then planktonic evolution is ruled by protection and not by competition. The many different shapes and life cycles reflect responses to specific attack systems.

Suppose that competition for light rather than protection were the driving force in shaping pelagic ecosystems. Faced with a single, optimal solution, algae could well have evolved more efficient photosynthetic machinery. Improved energy use would favour production of hydrocarbons as both a buoyancy aid and a reserve substance. The ocean surface would then be covered with oily scum that would, as well as changing the planetary heat budget, severely reduce evaporation and hence rainfall on the continents, where life as we know it could not then have evolved. Luckily for us, this did not happen, and we have our blue, white and brown planet with its smudges of green, instead of dark green (or even black) oceans and bare, brown continents. ■

Victor Smetacek is at the Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany.

FURTHER READING

Margulis, L., Matthews, C. & Haselton, A. (eds) *Environmental Evolution* 2nd edn (MIT Press, Cambridge, MA, 2000).
Dawkins, R. *The Extended Phenotype* (Oxford Univ. Press, Oxford, 1982).
Raven, J. A. *Limnol. Oceanogr.* **42,** 198–205 (1997).
Verity, P. G. & Smetacek, V. *Mar. Ecol. Prog. Ser.* **130,** 277–293 (1996).

© 2001 Macmillan Magazines Ltd

Electronics in a spin

Michael L. Roukes

Electrons can be shuttled around according to their 'spin', as well as their charge. Do this efficiently in a semiconductor and the future of electronics starts to look very different.

Essentially all semiconductor technology is based on electronic devices, such as transistors, that operate by internally shuttling small packets of electronic charge. In the past decade, however, devices have been developed that exploit the electron's 'spin' rather than its charge[1]. As a consequence of quantum mechanics, all electrons have spin, which gives them a magnetic moment — much as if they were tiny bar magnets. Many advances have been made in this new field of spin electronics, known also as spintronics, but scientists have not yet reached the same exquisite control over the flow of spin in microscale devices as they have over the flow of charge in conventional electronic devices.

On page 770 of this issue, Malajovich *et al.*[2] report an important advance in semiconductor spintronics. They show that it is feasible to transfer spins coherently (in a spin-aligned state), and with high efficiency, across interfaces between different semiconductor materials. This achievement is a milestone in the quest to build devices that exploit electron spin.

One of the most compelling and intriguing applications of semiconductor spintronics is quantum computation. Quantum computers could be enormously powerful, in part because binary digits or 'bits' in classical computers have just two available states (0 or 1), whereas quantum bits or 'qubits' have many more, which are all possible complex mixtures of 0 and 1. Robust examples of these elemental qubits have, so far, been demonstrated only in the form of ultracold atoms or ions. Such qubits are typically shrouded deep within a forest of specialized equipment that maintains them in an ultra-high vacuum, where they are nurtured and interrogated by a vast and intricate assemblage of precision laser instrumentation. But practical quantum computers will ultimately require arrays, comprising many qubits, all operating cooperatively and coherently. Scaling up the existing technology to allow the trapping of many ultracold atoms and ions at the requisite level of complexity is clearly impractical. The success of microelectronics, however, raises high expectations that electron or nuclear spins within semiconductor devices may provide the key to such scalability. But spins in semiconductors must possess several attributes before they can be harnessed to build usable qubits.

Work in spintronics has emerged from research into spin behaviour in metallic devices. Metal-based spintronics is already being used in state-of-the art computer hard drives and other magnetic devices. But spintronics is now becoming increasingly centred on semiconductor materials and processing techniques. Innovations in the mainstream semiconductor electronics industry have advanced these both to the point that a truly remarkable degree of material purity and technical precision can be attained. These innovations provide an opportunity to assemble new devices and materials in ways that are compatible with existing electronics methodology. In addition to potential applications in quantum information systems, spintronics may also lead in the near future to the development of semiconductor devices capable of performing high-speed logic and memory operations, similar to that achieved by conventional charge-based electronics, but at a fraction of the power.

At a fundamental level, nuclear and electronic spins are responsible for a material's magnetism. So the ability to sense and control magnetic properties is clearly at the heart of spintronic device operation. There are three long-standing prerequisites for success in semiconductor spintronics, which involve three fundamental aspects of spintronic device operation. First, all spintronic devices rely on spin-polarized carriers: electrons that are predominantly in a spin-aligned state, being nearly all 'spin-up' or 'spin-down'. Robust spin polarization arises naturally in ferromagnetic materials, which can serve as a stable source of spin-polarized carriers under ordinary conditions. Second,

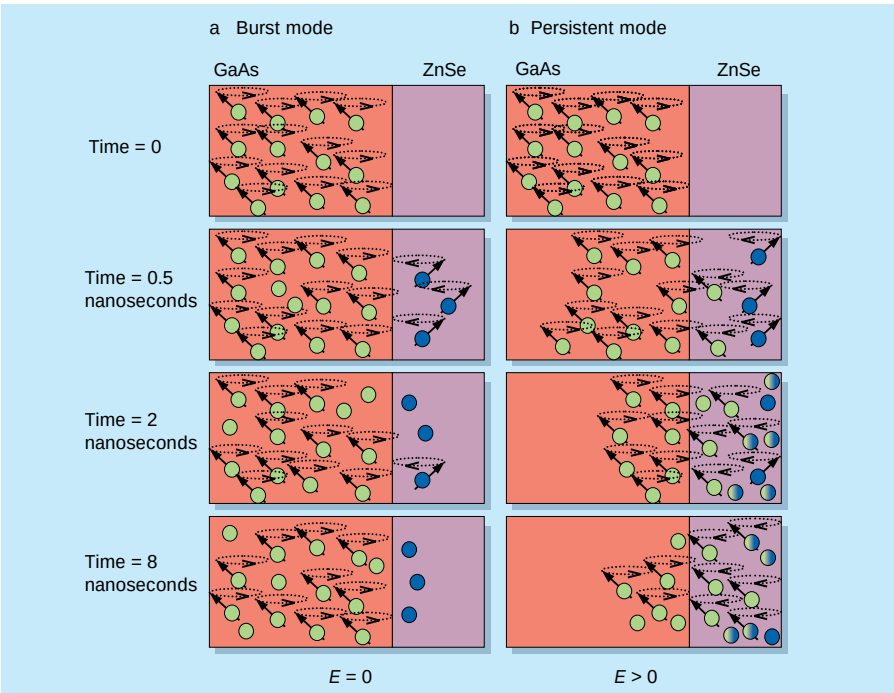


Figure 1 **Electron spin transfer across an interface between two different semiconductors (gallium arsenide, GaAs, and zinc selenide, ZnSe) in a magnetic field.** a, **Burst mode. When no external electrical field is applied ($E = 0$), only a short, fast burst of spins (circles) crosses the interface. Most of the spins remain trapped within the GaAs spin reservoir. Once spins cross into ZnSe they actually precess in the opposite direction and decay at a rate characteristic of ZnSe spins (blue).** b, **Persistent mode. When an electrical field ($E > 0$) is used to drag spins from the GaAs reservoir into the ZnSe layer, spins that have just arrived into ZnSe carry with them information about the spin precession in the reservoir (green). Spins that have been in ZnSe for a while (green/blue) have lost that information, and spins that crossed in the initial burst (blue) never had it, so these two spin species do not contribute to the net spin. Meanwhile, the continuous supply of fresh spins from GaAs results in a spin current, which is not characteristic of ZnSe, but evolves in the fashion of GaAs spins.**

© 2001 Macmillan Magazines Ltd

the mobile spins need to be transferred efficiently across interfaces between different semiconductor materials. Because such interfaces are ubiquitous in complex electronic devices, this transfer must proceed without appreciable loss of spin polarization. Third, once the spins are transferred into the heart of a device, where the switching or memory function actually occurs, their polarization must be preserved long enough for the desired operation to be carried out.

Progress towards satisfying these prerequisites has been made just within the past few years. Semiconductors that are not normally magnetic, such as gallium arsenide, have been turned into ferromagnetic materials by adding magnetic dopants[3–5]. Some of these new materials remain ferromagnetic even at relatively high temperatures. Crucially, their compatibility with conventional non-magnetic semiconductors enables optimal interfaces to be constructed in all-semiconductor spintronic devices. This allows spin-polarized electrons to be transferred efficiently between the respective material layers. This has clearly been demonstrated by the recent success of spintronics-based light-emitting diodes, which use electrical spin injection to emit circularly polarized light[6–8]. Optical techniques have been used to excite spins in doped semiconductors. In a magnetic field these spins continuously precess (rotate) about the field. The important point is that this precession has been observed to persist coherently for surprisingly long periods of time[9].

These are promising starts, yet not all of the pieces are in place. For example, efficient electrical injection (transfer) of spins across interfaces has, so far, only been demonstrated in structures involving semi-magnetic materials[6,8], which provide spin-polarized electrons solely when subjected to very high magnetic fields, that is, at levels available only in the laboratory from superconducting magnets. By contrast, when electrons from a ferromagnetic superconductor, which have intrinsic spin polarization even in the absence of an external field, are electrically injected from a ferromagnetic semiconductor across an interface into another non-magnetic semiconductor, the results to date are much less pronounced. Although the demonstration under these conditions is an important achievement in itself, the efficiency attained so far is only a few per cent[7].

Malajovich and co-workers[2] have now shown that spin-transfer efficiency can be increased by up to a factor of 40 by using electric fields to 'drag' spins across a semiconductor interface. The authors first use a laser pulse at the right frequency to create a long-lived reservoir of spin-polarized carriers, here in gallium arsenide, and then, in two separate experiments, show that both externally applied and built-in electric fields can be highly effective at transferring spins from this reservoir to another semiconductor, here zinc selenide (Fig. 1). In the latter case, internal built-in electric fields are created by the explicit inclusion of electrical dopants of different polarities in the semiconductor layers. In this way, a 'natural' potential arises at the interface between the differently doped semiconductors; this forms the basis for semiconductor diodes.

Electrons in solids have different spin properties to those in a vacuum. The electron's magnetic properties (and therefore its spin) are quantified in terms of a '*g*-factor', which characterizes the electron's energy in a magnetic field. One of the intriguing aspects of the new work is that, under the best spin-transfer conditions, the entire population of transferred spins (within the collecting layer) is found to behave, overall, with the character of their original environment — in this case the original semiconductor reservoir.

This occurs because of a newly identified, persistent mode of spin transfer, which makes the reservoir act, in effect, as a spin 'battery'. This means that, under special conditions, a spin reservoir can be made to continuously source a spin-polarized current across the interface until the reservoir itself is depleted.

This is demonstrated by Malajovich *et al.*[2] using a sensitive technique that allows them to interrogate the identity (effective *g*-factor) of the ensemble of transferred spins. They find that the effective *g*-factor for the population of transferred spins can be continuously metamorphosed from a value characteristic of the spin reservoir to that of the collecting semiconductor. The precise behaviour depends on the strength of the effective electric field. This unexpected phenomenon may provide the basis for what the authors call multifunctional spintronic devices, in which the character of the spin currents can be controlled by both electric and magnetic fields. This prospect injects an intriguing new 'spin' into the emerging technology of semiconductor spintronics. ■

Michael L. Roukes is in the Department of Physics, California Institute of Technology 114-36, Pasadena, California 91125, USA.
e-mail: roukes@caltech.edu

1. Prinz, G. A. *Science* **282,** 1660–1663 (1998).
2. Malajovich, I., Berry, J. J., Samarth, N. & Awschalom, D. D. *Nature* **411,** 770–772 (2001).
3. Ohno, H. *et al. Phys. Rev. Lett.* **68,** 2664–2667 (1992).
4. Munekata, H. *et al. Appl. Phys. Lett.* **63,** 2929–2931 (1993).
5. Ohno, H. *Science* **281,** 951–956 (1998).
6. Fiederling, R. *et al. Nature* **402,** 787–790 (1999).
7. Ohno, Y. *et al. Nature* **402,** 790–792 (1999).
8. Jonker, B. T. *et al. Phys. Rev. B* **62,** 8180–8183 (2000).
9. Kikkawa, J. M. & Awschalom, D. D. *Phys. Today* **52,** 33–39 (1999).

Plant microbiology

Quieting the raucous crowd

Jared R. Leadbetter

Many bacteria communicate by using dedicated signalling molecules. Signal-degrading enzymes from other bacteria interrupt the conversation, and can protect tobacco and potato plants from infection.

Most of us have had occasion to reach into a refrigerator crisper drawer to sponge out the remnants of a wilted, slimy head of lettuce. One of the major culprits of this horror is *Erwinia*, a bacterium that also causes economically important wilts and soft-rots of crops. On page 813 of this issue, Dong and colleagues[1] introduce a promising new way of controlling *Erwinia* infections — an approach that might also have implications for treating some human diseases.

The past 30 years have seen the discovery that *Erwinia* and many other bacteria can not only sense, but also respond to, increases in their population density. When crowded, they dramatically change their physiology. This process, called quorum sensing, is often associated with the bacteria shifting from a free-living lifestyle to becoming dependent on a particular host[2]. Many of these bacteria produce, monitor and respond to the accumulation of soluble signalling molecules — quorum-sensing signals. Molecules in one class of these signals have a specific chemical structure described as an acyl-homoserine lactone (AHL; Fig. 1). In the case of *Erwinia*, AHLs activate the expression of diverse 'virulence factors' such as slime, and enzymes that solubilize polysaccharides[3,4] to provide the bacteria with nutrients.

Dong *et al.*[1] took advantage of *Erwinia*'s quorum-sensing signals to develop a new way of controlling quorum-regulated processes. They call their technique 'quorum quenching'. Instead of trying to target bacterial growth or viability directly, the authors focused on ways to degrade the AHL signals as they are produced, with the aim of reducing or eliminating the expression of quorum-regulated virulence factors. To do this, Dong *et al.* constructed transgenic tobacco and potato plants expressing the bacterial gene *aiiA*. This gene encodes an enzyme that renders AHLs biologically inactive. Remarkably, the transgenic plants became resistant to *Erwinia* infection, even when unrealistically high numbers of *Erwinia* cells were introduced into wounded tissues. In rare cases where there were localized signs of disease, the transgenic plants managed to fight off the infection over time. By comparison, control plants succumbed to an uncontrol-

© 2001 Macmillan Magazines Ltd

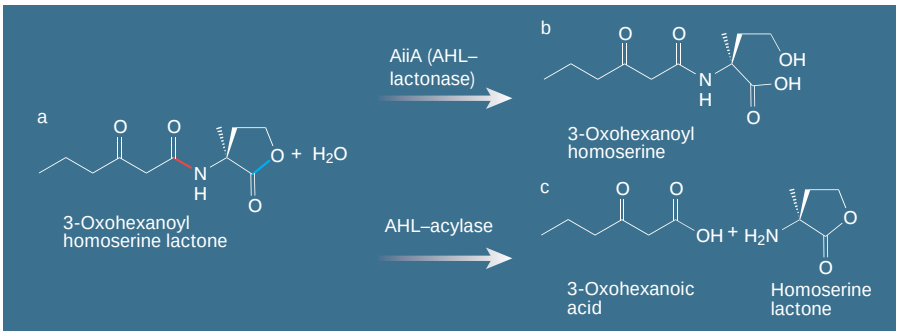


Figure 1 **Disabling soft rot: biochemical routes to the inactivation of acyl-homoserine lactone (AHL) molecules.** a, **The AHL *N*-3-oxohexanoyl-L-homoserine lactone is a quorum-sensing signal used by the *Erwinia* bacteria, which cause soft rot in certain crops, to control their ability to cause disease in plants. The bonds that can be broken by two inactivating enzymes are shown in blue and red.** b, **Dong *et al.*[1] have used the protein AiiA to hydrolyse the blue bond and generate the inactivated signal *N*-3-oxohexanoyl-L-homoserine.** c, **3-Oxohexanoic acid and L-homoserine lactone are the products of an uncharacterized AHL-acylase found in the bacterium *Variovorax paradoxus*.**

lable infection, even when challenged with comparatively low numbers of bacteria.

The result is a triumph for the approach of mining bacterial metabolic diversity for applied uses. In this case, Dong *et al.* capitalized on the discovery that many bacteria can degrade, and so inactivate, AHLs. Earlier, the same group[5] had isolated hundreds of bacterial species from soil and screened them for the ability to degrade AHLs. From this collection, a bacterium of the genus *Bacillus* was identified as being especially promising. It was from this strain that *aiiA* was cloned. As well as their exciting observations of plants expressing this gene, the authors[1] also show how the AiiA protein inactivates AHL signals: it degrades the molecule's lactone ring, so acting as an AHL-lactonase (Fig. 1).

Another soil bacterium has been identified that also degrades these signals, through an uncharacterized AHL-acylase activity[6]. Clearly, there are different mechanisms for the biodegradation of AHLs, inviting comparisons to the way in which penicillins can be inactivated.

Enzymes that degrade AHLs might be useful in controlling many other quorum-sensing bacterial populations, such as biofilms. These are dense populations of bacteria, attached to some surface, that grow to form complex, tissue-like structures; they are extremely resistant to antibiotics and have been implicated in many human diseases[7]. For example, the formation of biofilms by *Pseudomonas aeruginosa* is common in the lungs of people with cystic fibrosis. Here, quorum sensing regulates the expression of virulence factors and the development of biofilms[8]. As aerosols of mucous-degrading enzymes and other chemicals are already used to treat lung disease in cystic fibrosis patients, it might prove fruitful to add AHL-degrading enzymes to the mix.

But in the rush to develop products based on AHL-lactonases that will no doubt come, microbiologists should not lose sight of the significance of AHL-degrading enzymes to the physiology and ecology of the bacteria that produce them. There are questions to be answered, such as: do these enzymes have narrow or broad substrate specificity? And how are they regulated in the cell?

Moreover, what is the effect of signal degradation on the cells involved, and in the environment at large? No doubt bacteria have a stake in disarming the quorum-sensing systems of other species. For example, *Erwinia* regulates its synthesis of carbapenem, an antibacterial compound, through AHL-dependent mechanisms. Presumably, carbapenem keeps other bacteria away from the nutrient cache accumulated during plant infection. Those other species may use signal degradation to circumvent such a barrier. As revealed by genome sequencing, nearly identical counterparts of *aiiA* are present in several *Bacillus* species. Curiously, though, the quorum-sensing plant pathogen *Agrobacterium tumefaciens* may also have an *aiiA* relative, called *attM*, which is associated with genetic loci that encode proteins needed for this bacterium to colonize plant roots[9]. It is not yet clear why *Agrobacterium*, which makes and responds to AHL signals, might also need to degrade them.

Studies of the complex competitive interactions among microbes will no doubt be very informative. One epic story of microbial rivalry encompasses counter-evolutionary tales of naturally occurring β-lactam antibiotics, the enzymes that degrade them (β-lactamases) and β-lactamase inhibitors, such as clavulanic acid. In a parallel to this narrative, bacteria that make AHLs may have evolved resistance to the enzymes that break the signals. Some bacteria might produce factors that bind and block the active sites of AHL-degrading enzymes. Others might have evolved enzymes that recycle the signals after hydrolysis of their lactone rings. Nature throws punches and counter-punches: it will surely be both interesting and useful to investigate and anticipate them. In the meantime, Dong *et al.*[1] have returned *Erwinia*'s first blow. In doing so, they have revealed a soft spot in soft-rot diseases, perhaps of a type to be found in other quorum-sensing pests and pathogens. ■

Jared R. Leadbetter is in the Program of Environmental Science and Engineering, California Institute of Technology, Pasadena, California 91125-7800, USA.
e-mail: jleadbetter@caltech.edu

1. Dong, Y.-H. *et al. Nature* **411,** 813–817 (2001).
2. Fuqua, C., Winans, S. C. & Greenberg, E. P. *Annu. Rev. Microbiol.* **50,** 727–751 (1996).
3. von Bodman, S. B. & Farrand, S. K. *J. Bacteriol.* **177,** 5000–5008 (1995).
4. Swift, S. *et al. Mol. Microbiol.* **10,** 511–520 (1993).
5. Dong, Y.-H., Xu, J. L., Li, X. Z. & Zhang, L. H. *Proc. Natl Acad. Sci. USA* **97,** 3526–3531 (2000).
6. Leadbetter, J. R. & Greenberg, E. P. *J. Bacteriol.* **182,** 6921–6926 (2000).
7. Donlan, R. M. *Emerg. Infect. Dis.* **7,** 277–281 (2001).
8. Davies, D. G. *et al. Science* **280,** 295–298 (1998).
9. Matthysse, A. G., Yarnall, H., Boles, S. B. & McMahan, S. *Biochim. Biophys. Acta* **1490,** 208–212 (2000).

Carbon cycle

The roots of the matter

F. Stuart Chapin III and Roger W. Ruess

Perhaps the most scientifically challenging phase of the terrestrial carbon cycle occurs below ground. Innovative experiments, carried out in northern Sweden, illustrate the huge influence of roots and associated fungi.

Photosynthesis by terrestrial vegetation accounts for about half of the carbon that annually cycles between Earth and the atmosphere[1]. Although above-ground plant production is relatively well documented from field measurements and globally distributed satellite observations, the quantity of carbon that plants transfer below ground is not well known. Microvideo cameras provide direct observations of the growth, longevity and decomposition of fine roots[2,3]. However, other large components of below-ground plant production, such as exudation of organic compounds by roots and transfer of carbohydrates to associated mycorrhizal fungi, are difficult to study non-destructively and have not been estimated. Estimates of the contribution of root respiration to total CO_2 efflux from soil range from 10% to 90%, with methodological uncertainties accounting for most of this variation[4].

On page 789 of this issue, Högberg and colleagues[5] present a new approach that

© 2001 Macmillan Magazines Ltd

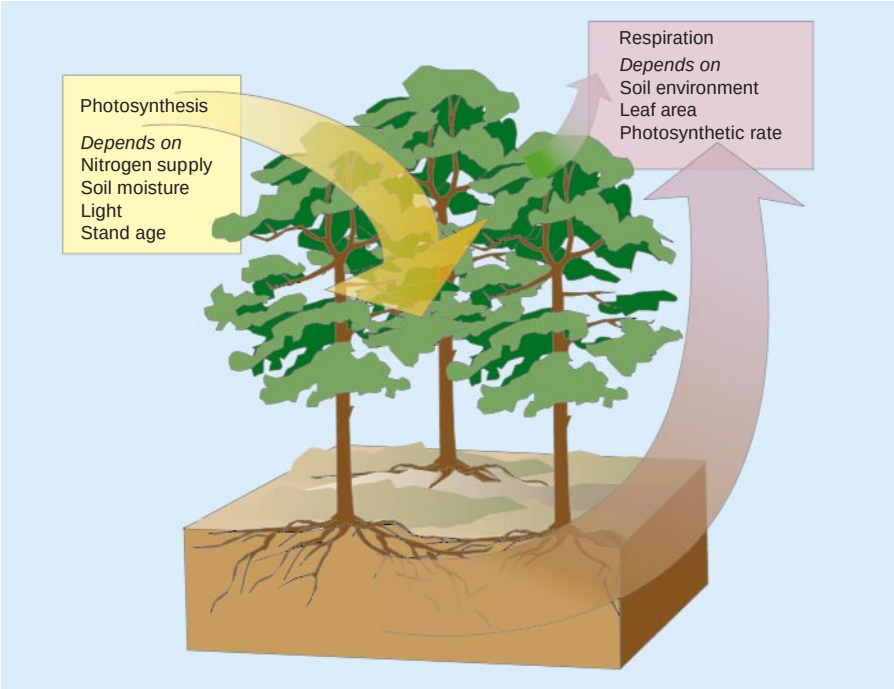


Figure 1 **The main carbon fluxes and their controls in boreal forests. The largest fluxes are carbon input from photosynthesis and carbon loss in respiration from roots, mycorrhizal fungi and other root-associated microorganisms. More minor loss is due to leaf and trunk respiration, as indicated by the small arrow. The boxes show the various factors that control the fluxes. As the results of Högberg *et al.*[5] show, events above ground are partly controlled by events below ground, and vice versa, meaning that all parts of a plant are affected by environmental change.**

provides an integrated estimate of the proportion of soil respiration derived from roots and their symbiotic mycorrhizae. Their experiments involved 'girdling' all trees in experimental plots of Scots pines in northern Sweden (that is, in boreal forest). In girdling, bark 1.5 m above the ground was stripped to a depth that allowed upward water flux but prevented downward transport of carbohydrates to roots and mycorrhizae. Soil respiration was measured as the rate of CO_2 accumulation in closed chambers, 24 cm in diameter, inserted into the soil surface. By comparing soil respiration in girdled plots, in which root and mycorrhizal respiration had been eliminated, with control plots where they had not, Högberg *et al.* concluded that over half of the soil respiration in the forest derived from the current year's photosynthesis.

As the authors point out, their estimate of below-ground plant respiration is conservative because root carbohydrate reserves were probably mobilized to support root function in the absence of a supply from above. The soil respiration measured in treated plots may therefore have included some respiration of tree roots, as well as the respiration of roots of understorey shrubs that were not manipulated. In addition, increased root death may have stimulated greater respiration levels of heterotrophic (decomposer) organisms than those that would naturally occur in forests. All of these sources of error would lead to an underestimate of the proportion of soil respiration accounted for by roots.

The difference between carbon released in soil respiration and that in above-ground leaf litter has been used as an index of total below-ground carbon allocation in forests[6]. These indices suggest that, relative to temperate forests, boreal vegetation allocates a higher proportion of fixed carbon to below-ground structures[7]. Högberg *et al.* now show that at least half of the soil respiration in these systems is derived from roots and mycorrhizae. Their approach minimizes artefacts associated with smaller-scale trenching experiments, which frequently alter soil moisture and the structural integrity of roots.

The ratio of plant-associated respiration derived from roots to that from their symbiotic mycorrhizae remains a mystery. Direct observations of fine roots in boreal forests point to rapid rates of growth and decomposition relative to above-ground tissues[3]. But there are few data on the proportion of plant production in this or any other terrestrial ecosystem that is allocated to mycorrhizae. Consequently, the current measures of global terrestrial production may be serious underestimates of the true values.

Högberg and colleagues' demonstration[5] of a rapid decline in soil respiration after girdling shows the tight integration of the whole-plant response to environmental changes (Fig. 1). These results are consistent with studies showing that roots and soil respiration respond to above-ground herbivory[7],

NATURE

100 YEARS AGO

Hailstorm Artillery. **In the absence of any recognised English equivalent for the expressive German term *Das Wetterschiessen*, I have thought it best in the heading of this article to avoid a literal translation of it lest it should give rise to misunderstanding. 'Weather shooting' does not refer to any haphazard or empirical attempts to foretell the weather, but to a practice which has lately come to have great vogue in Styria, Italy and elsewhere of firing off charges of gunpowder to protect the vineyards against injury from hail. So popular indeed has the practice become in some districts that there is danger of the cost of protection exceeding that of any damage likely to be caused by the hail. The idea that the weather can be affected by the discharge of gunpowder is not a new one. There have been various traditions of rain falling after, and presumably in consequence of, the cannonade of a battle… Weather shooting as now practised has, however, a more definite purpose than merely causing rain. Its object is to prevent the downpour of hail by shooting when thunder or hail clouds threaten. The recent development, which has spread very widely, is most conspicuously represented by the arrangements of Bürgermeister Stiger, of Windisch-Feistritz, in Styria, where they were originally introduced in 1896 in the form of a vine-dressers' volunteer artillery.**
From *Nature* 13 June 1901.

50 YEARS AGO

Profile of Science. **Setting out unashamedly to interest those who would not touch a 'scientific' book, Ritchie Calder must be congratulated on the adventurous nature of his approach. He is aware that most people are interested in the biographical details of famous men and uses this as the way to introduce the four main subjects of the atom, radar, penicillin and vitamins. Each topic is complete in itself, and is portrayed in relation to its general scientific background; the story of penicillin, for example, is cunningly interspersed with accounts of the invention of the microscope, the discovery of germs and chemical therapeutics as well as up-to-date information about more recent investigations with antibiotics like streptomycin, chloromycetin and aureomycin… Perhaps the only criticism that can be made of his excellent impressionistic book is that occasionally the popular style becomes a little too 'hail-fellow-well-met' for those with more retiring natures.**
From *Nature* 16 June 1951.

© 2001 Macmillan Magazines Ltd

availability of nutrients[8,9] and light[10], and other factors that govern plant carbon gain. However, roots also respond directly to the below-ground environment. For instance, root respiration increases exponentially with increases in soil temperature[11], which could explain why fine-root elongation[3,12] and root respiration[5] in boreal regions peak in mid-to-late summer, when soil temperatures are warmest. The interaction of demand for carbohydrates, as determined by the root environment, and the above-ground capacity to supply carbohydrates, as determined by photosynthesis, together govern the below-ground carbon flux and therefore root growth and respiration. Conversely, photosynthetic rate is strongly influenced both by leaf environment (light and temperature) and by soil resources (nutrients and water). This principle — that each physiological process responds to the plant's total environment — may seem obvious. In many ecosystem models, however, root respiration is often estimated as a simple function of soil temperature and total root biomass, and photosynthesis as a function of air temperature and light.

Högberg and colleagues' results, which show a high proportion of carbon allocation below ground and tight integration among all plant processes, suggest that cycling of carbon through terrestrial ecosystems is both larger in magnitude and more complex than was previously appreciated. Clearly, above-ground production is only the tip of the terrestrial carbon-cycle iceberg. ■

F. Stuart Chapin III and Roger W. Ruess are at the Institute of Arctic Biology, University of Alaska, Fairbanks, Alaska 99775, USA.
e-mails: fschapin@lter.uaf.edu
ffrwr@uaf.edu

1. Schimel, D. S. *Glob. Change Biol.* **1,** 77–91 (1995).
2. Hendrick, R. L. & Pregitzer, K. S. *Ecology* **73,** 1094–1104 (1992).
3. Ruess, R. W., Hendrick, R. L. & Bryant, J. P. *Ecology* **79,** 2706–2720 (1998).
4. Hanson, P. J., Edwards, N. T., Garten, C. T. & Andrews, J. A. *Biogeochemistry* **48,** 115–146 (2000).
5. Högberg, P. *et al. Nature* **411,** 789–792 (2001).
6. Raich, J. W. & Nadelhoffer, K. J. *Ecology* **70,** 1346–1354 (1989).
7. Ruess, R. W., Van Cleve, K., Yarie, J. & Viereck, L. A. *Can. J. For. Res.* **26,** 1326–1336 (1996).
8. Nadelhoffer, K. J. *New Phytol.* **147,** 131–139 (2000).
9. Burton, A. J., Pregitzer, K. S. & Hendrick, R. L. *Oecologia* **125,** 389–399 (2000).
10. Craine, J. M., Wedin, D. A. & Chapin, F. S. III *Plant Soil* **207,** 77–86 (1999).
11. Burton, A. J., Pregitzer, K. S., Zogg, G. P. & Zak, D. R. *Can. J. For. Res.* **26,** 1761–1768 (1996).
12. Tryon, P. R. & Chapin, F. S. III *Can. J. For. Res.* **13,** 827–833 (1983).

Animal behaviour

Homing in on ant navigation

Mandyam V. Srinivasan

The ability of ants to travel far from their nests in search of food and to accurately chart their way back has intrigued researchers. Presenting the insects with a hilly challenge now throws light on these navigational skills.

Foraging animals have to be able to find their way to a food source and back to their home or refuge. The Saharan desert ant, *Cataglyphis fortis*, is an especially impressive navigator, unerringly traversing hundreds of metres across barren landscapes that are almost completely devoid of landmarks. That's why Wohlgemuth, Ronacher and Wehner (page 795 of this issue)[1] used trained individuals of this species to investigate one aspect of navigation: how the ants calculate the distance they have travelled in their equivalent of a car's odometer.

Saharan desert ants (Fig. 1) are long-legged, elegant creatures. In navigating, they rely on an accurate path-integration system, a mechanism that continually keeps track of where the animal is in relation to home. To perform path integration, the animal must monitor its forward motions and turns, and sum them up to establish its current position in relation to home. Mathematically, this amounts to subdividing the outbound trajectory into a number of small, straight segments, representing each segment by a vector of the appropriate length and direction, and summing these elementary vectors to obtain a 'homing' vector that specifies the distance and direction of home.

Exactly how the ant measures its forward motions and turns and puts them together is not yet clear, although there are some interesting possibilities. For instance, Müller and Wehner[2] found evidence that ants use quick and dirty approximations to compute the vector sums. Now Wohlgemuth *et al.* have illuminated an intriguing property of the ant's path-integration system that is revealed when these creatures are made to forage over undulating terrain.

In one series of experiments, ants were trained to forage at a feeding site by traversing 'hilly' terrain, simulated by a series of channels sloping alternately upward and downward in a sawtooth-like profile (see the cover of this issue). When these trained ants were tested on flat terrain, simulated by a horizontal channel, they searched for the food after they had walked a considerably shorter distance than in the training. Conversely, when ants were trained to find food at a fixed distance in flat terrain and were then tested in hilly terrain, they walked a longer distance before commencing their search. In each test, the ants searched for the food at a location that corresponded not to the actual distance walked in the training, but to the projection of this distance on the horizontal plane.

The authors conclude that these desert ants do not navigate by measuring the distance they actually walk on the undulating terrain. Rather, they project their movements onto a fictitious horizontal plane and register their position relative to the nest on this plane. The advantage of this strategy is that the location of the nest (or a bountiful supply of food) can be pinpointed accurately irrespective of the undulations of the ground. So navigation can be accurate even when the return route is different from the outbound route and has a different profile, as is commonly the case in natural foraging.

This is an attractive hypothesis. But the data do not exclude another possibility — that the ants navigate by performing path integration in all three dimensions, rather

R. WEHNER

Figure 1 **Impressive navigator — the Saharan desert ant, *Cataglyphis fortis*.**

© 2001 Macmillan Magazines Ltd

than projecting their movements onto two. The behaviour of the ants in each test can then be explained on the basis that they are searching for the food at the same location, in three-dimensional space, as in the training. One way to distinguish between the two possibilities might be to train ants to find food by climbing up to the end of a ramp, and then to test them in an L-shaped channel composed of a horizontal arm followed by a vertical arm that leads to the trained position of the food. If the ants are projecting their motions onto the horizontal plane, they should search at the end of the horizontal arm. On the other hand, if they are performing path integration in all three dimensions, they should climb up the vertical arm and search at its end.

Fully three-dimensional path integration would be of obvious value to ant species that inhabit arboreal environments. Its relevance to desert ants remains to be assessed, but Wohlgemuth *et al.* say that they plan to do just that. ■

Mandyam V. Srinivasan is at the Centre for Visual Science, Research School of Biological Sciences, PO Box 475, Australian National University, Canberra, ACT 2601, Australia.
e-mail: m.v.srinivasan@rsbs.anu.edu.au

1. Wohlgemuth, S., Ronacher, B. & Wehner, R. *Nature* **411,** 795–798 (2001).
2. Müller, M. & Wehner, R. *Proc. Natl Acad. Sci. USA* **85,** 5287–5290 (1988).

Biomineralization

Crystals, asymmetry and life

Lia Addadi and Steve Weiner

Understanding the formation of asymmetrical shapes during the growth of symmetrical crystalline structures is a first step towards understanding asymmetry in biology.

Lack of symmetry is a striking property of the biological world. It is seen at the molecular, macromolecular, cellular and tissue level. The induction of molecular asymmetry has been widely studied, but the transfer of asymmetry at the molecular level into asymmetry at the macroscopic level — over distances of nanometres, micrometres and longer — is still poorly understood.

The coiled shell of the foraminifer *Globigerina pachyderma*, a marine microorganism, provides a remarkable example of macroscopic asymmetry (Fig. 1a). In the Arctic and Antarctic, more than 98% of these calcium carbonate shells coil in a counterclockwise direction, but in temperate and tropical regions, more than 98% of them coil in the clockwise direction[1]. This is a simple illustration of chirality — the two forms of shell are non-superimposable mirror images of each other.

There are many examples of chiral shapes in biologically formed minerals, and a lot can be learned about how they are created by studying how chiral information is transferred from molecules to crystalline surfaces. On page 775 of this issue, Orme *et al.*[2] examine the way in which molecules of aspartic acid bind to growing calcite crystals (a form of calcium carbonate). Calcite is not chiral, whereas the two molecular forms of aspartic acid are mirror images of each other. Interactions between chiral aspartic acid and calcite causes macroscopic hillocks to form on the crystal surface. These hillocks have shapes that are also mirror images of each other (Fig. 1b).

The changes in surface morphology reflect the chirality of the aspartic acid molecule. The same process occurs when the crystal is partially dissolved in the presence of aspartic acid, leaving behind pits that are mirror images of each other. Honess first observed this dissolution effect more than 60 years ago[3], but the molecular explanation was missing. The paper by Orme *et al.* now fills in this basic gap in our understanding.

In 1894, Pierre Curie laid the conceptual foundation for this work by stating[4]: "When certain causes produce certain effects, the elements of symmetry of the causes must be found in the effects produced. When certain effects reveal a certain asymmetry, this asymmetry must be found in the causes that originated them. The opposite is not necessarily true… namely the effects produced may be more symmetrical than the causes that produced them." These statements imply that whenever chirality is seen in a macroscopic object, there must be a chiral component in its precursors. The contrary is not necessarily true because the interactions of a chiral molecule with a non-chiral object do not inevitably 'reduce' the symmetry of the non-chiral object.

In the system studied by Orme *et al.*[2], the reduction of calcite symmetry occurs at a crystal interface during growth (or dissolution). A detailed understanding of this process has emerged from a series of investigations on organic crystals[5,6]. Although crystal surfaces largely follow the arrangement of the molecules within the crystal bulk, they also have surface structures that are different from that of the bulk. Furthermore, if the growing surface is at an angle to the symmetry elements of the bulk structure, the symmetry that relates molecules at the surface is lower than the symmetry of the bulk. In

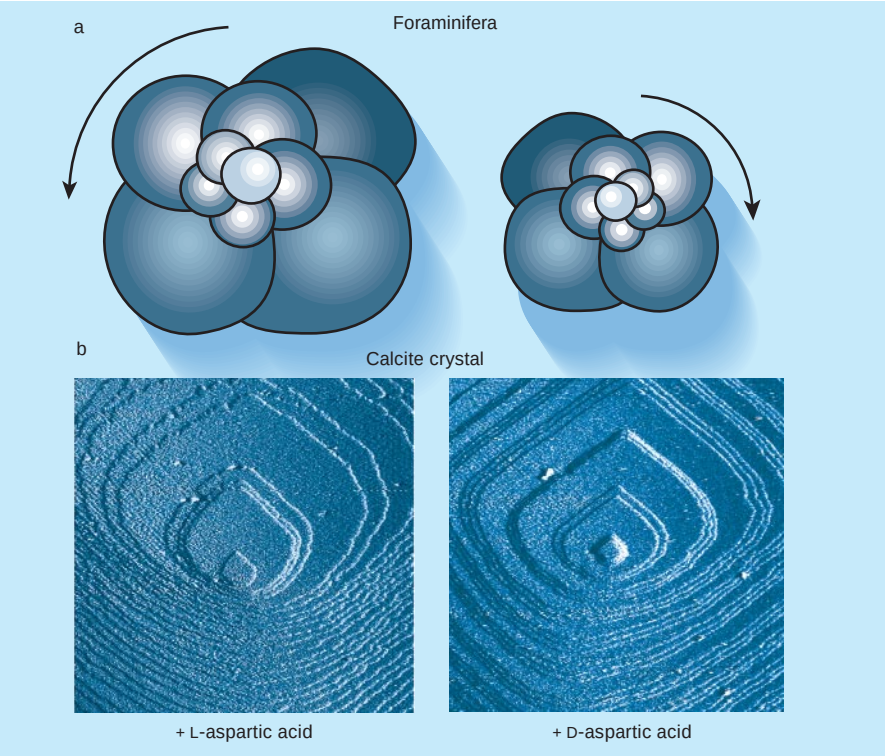


Figure 1 **Chiral shapes of biological and synthetic crystals.** a, **Laevorotatory and dextrorotatory shapes of specimens of the foraminifer *Globigerina pachyderma*[1], traced from micrographs.** b, **The hillocks seen in calcite crystals grown in the presence of L-aspartic acid or D-aspartic acid. Orme *et al.*[2] have studied how aspartic acid molecules bind to growing calcite crystals, and so have managed to explain how molecular chirality leads to macroscopic chirality.**

© 2001 Macmillan Magazines Ltd

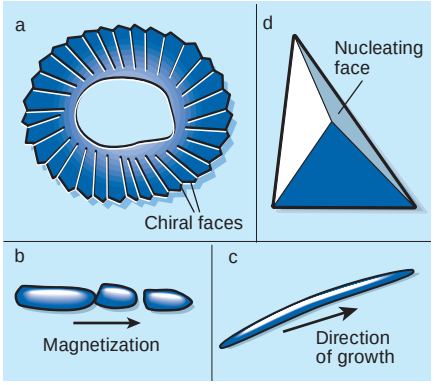


Figure 2 **Living organisms use macromolecules to control the growth of biological minerals — a process known as biomineralization. The following are examples of biologically produced single crystals with chiral morphology.** a, **The disc-shaped elements of a coccolith, displaying the chirality of the single crystals[8].** b, **A string of magnetite crystals from a magnetotactic bacterium. The typical bullet-shaped single crystals have chiral morphology, and are elongated along a specific crystallographic direction, which is the direction of magnetization[9].** c, **Curved spicule of the sponge *Sycon*. The spicule grows along one of three symmetry-related equivalent directions[10].** d, **Morphology of calcium oxalate crystal from the leaves of tobacco plants[11]. (Morphologies were traced from micrographs.)**

other words, the surface can be chiral, even if the bulk structure is not. Recently, Reeder explained the selective occlusion of impurities through specific growing of calcite crystal faces using this concept[7].

There are several examples of crystals formed by biological processes (biomineralization) that have macroscopic chirality with only one of the possible mirror-images being produced. Many *Coccolithophoridae*, a group of calcifying marine algae, cover their cell surfaces with disc-shaped structures[8]. Each disk consists of a radial array of single calcite crystals, each of which has macroscopic chirality (Fig. 2a). Bacteria that can sense magnetic fields form strings of single crystals of the magnetic mineral magnetite, each of which is enveloped in a membrane[9]. The magnetic field generated by the string of crystals is sufficient to align the bacterium in the Earth's magnetic field. The single crystals have macroscopic chirality (Fig. 2b).

The sponge *Sycon* has calcitic spicules, some of which grow into a sabre shape (Fig. 2c) along one of three possible symmetrically identical directions[10]. The plane of symmetry of the sabre-shaped spicule, however, does not coincide with any symmetry element of the *Sycon* structure. Thus the integration of the molecular symmetry with the macroscopic symmetry is chiral.

The calcium oxalate crystals from the leaves of tobacco or tomato plants have been known for more than a century to have chiral morphology, although their exact shape was only recently worked out[11] (Fig. 2d). Each plant species always forms crystals with the same morphology at the same tissue site, implying that the chirality is under genetic control.

A plausible explanation for asymmetry in the latter two examples is that the chiral information is contained in one or more biological molecules, and is transmitted to the final product during crystal nucleation from a unique structural plane. In essence these nucleating macromolecules distinguish between the front and back of the same crystal plane. So these biological examples offer a link between the highly complex processes involved in determining the helicity of the shell of *Globigerina pachyderma*, and the more manageable models studied in the laboratory, such as the growth of calcite in the presence of chiral amino acids.

In their study, Orme *et al.*[2] use atomic force microscopy and molecular modelling to show that when calcite is grown in the presence of either the L or D form of aspartic acid, the morphology of crystal steps and terraces becomes asymmetric. This change reflects the acid's chirality. They also calculate the free energy associated with the steps and used the known relation between step energy and rate of growth, namely that an unstable step grows quickly. They show that the energy minimum associated with some step edges disappears in the presence of aspartic acid, and a new minimum appears. The new position of the energy minimum depends on the manner in which either L or D aspartic acid binds to the step, and switches from one side of a crystal mirror symmetry element to the other, depending on the chirality of the additive. So the addition of D or L aspartic acid changes the rate of growth of the steps asymmetrically, and this results in turn in macroscopic chirality.

The study by Orme *et al.* helps to lay the foundation for one day understanding how the chiral morphology of coccoliths is induced, or why *Globigerina* coils in different directions, or why almost all gastropod shells coil in a clockwise direction. Perhaps information of this type will ultimately allow us to address fundamental issues of chirality in development, and the earliest processes that shape animals. ■

Lia Addadi and Steve Weiner are in the Department of Structural Biology, Weizmann Institute of Science, 76100 Rehovot, Israel.
e-mails: lia.addadi@weizmann.ac.il
steve.weiner@weizmann.ac.il

1. Bandy, O. J. *Paleontology* **34,** 671–681 (1960).
2. Orme, C. A. *et al. Nature* **411,** 775–779 (2001).
3. Honess, A. P. & Jones, J. R. *Bull. Geol. Soc. Amer.* **48,** 667–721 (1937).
4. Curie, P. *J. Phys (Paris)* **3,** 393 (1894).
5. Addadi, L., Berkovitch-Yellin, Z., Weissbuch, I., Lahav, M. & Leiserowitz, L. *Topics Stereochem.* **16,** 1–85 (1986).
6. McBride, J. M. & Bertman, S. B. *Angew. Chem. Int. Edn Engl.* **28,** 330–333 (1989).
7. Reeder, R. J. *Nature* **353,** 797–798 (1991).
8. Young, J. R., Davis, S. A., Bown, P. R. & Mann, S. *J. Struct. Biol.* **126,** 195–215 (1999).
9. Mann, S., Sparks, N. H. & Blakemore, R. P. *Proc. R. Soc. Lond. B* **231,** 477–487 (1987).
10. Aizenberg, J. *et al. Chem. Eur. J.* **7,** 414–422 (1995).
11. Bouropoulos, N., Weiner, S. & Addadi, L. *Chem. Eur. J.* **7,** 1881–1888 (2001).

Palaeontology

Australia's last giants

Jared M. Diamond

The huge creatures that once roamed Australia were wiped out at about the same time. From the latest analyses, these extinctions look more than ever to be a matter of not what but who dunnit.

During the Late Pleistocene and Holocene (the past 100,000 years or so), extinction waves of large animals swept across all of those parts of the world that had previously not been inhabited by humans[1]. The two largest such waves occurred in Australia and the Americas. Writing in last week's *Science*, Roberts *et al.*[2] describe how they have at last pinned down the date of the Australian wave, in a study rich with implications for the American wave.

Australia once supported 24 genera of megafauna (Fig. 1) — giant marsupials, reptiles and birds, including kangaroos 3 m high, marsupial equivalents of rhinoceroses and leopards, 1-ton carnivorous lizards like scaled-up Komodo dragons, and giant land-going crocodiles. At some point in the Late Pleistocene, 23 of these megafaunal genera became extinct, leaving behind only the *Macropus* kangaroos. Since the nineteenth century there has been controversy over the timing and cause of this event. Was it triggered by the arrival of human beings on the continent, or by changing climate — for instance, the increasingly arid conditions at the Last Glacial Maximum (19,000–23,000 years ago)?

Resolution of the question has been impeded by a lack of reliable dates. Both the arrival of humans and the extinction of megafauna in Australia lie just beyond radiocarbon dating's practical limit (around 40,000 years), so establishing the dates of both events has remained tantalizingly beyond our grasp. There are claims that megafauna survived into the Holocene (the past 10,000 years or so). But one reanalysis[3]

© 2001 Macmillan Magazines Ltd

AUSTRALIAN MUSEUM, NATURE FOCUS

Figure 1 **Megabite — one element of Australia's ancient megafauna, the giant goanna (*Megalania prisca*), dines off another, the marsupial *Diprotodon opatum*.**

has showed that all megafaunal ^{14}C dates younger than 28,000 years were invalidated either by doubtful association of dated samples with megafaunal bones themselves or by difficulties of ^{14}C dating near its limit.

For the problem of human arrival in Australia, these difficulties have been resolved by two dating techniques that reach beyond the ^{14}C limit. First, optical luminescence dating of sediment grains around bones or human artefacts yields the date when those grains were last exposed to sunlight — and hence the date when the bones or artefacts became buried. Second, $^{230}Th/^{234}U$ dating of the 'flowstones' enclosing the bones or artefacts yields those stones' crystallization age. Both methods give dates agreeing with ^{14}C dates within the useful time span of the ^{14}C method. Beyond that ^{14}C limit, they produce dates that agree with each other and are in the correct stratigraphic sequence.

Application of these two methods to human artefacts gave 56,000 ± 4,000 years ago as the oldest dates of artefacts, and hence the date of human arrival[4]. Roberts *et al.*[2] have now applied the methods to the youngest megafaunal remains (and hence to the date of megafaunal extinction), in a collaboration of 11 experts in Australian palaeontology and archaeology, and in Pleistocene dating. The team took great pains to find sites with major parts of intact megafaunal skeletons in a primary depositional context — that is, still lying in the sediments in which the animal died or first came to rest — as opposed to detached individual bones likely to have become washed into a sediment differing in age from the bones.

A previous study[5] of eggshells of the extinct giant bird *Genyornis newtoni* had estimated 50,000 ± 5,000 years ago as that species' youngest dates. Roberts *et al.* examined instead the burial ages of giant marsupials and reptiles at 28 sites selected on three criteria besides evidence of primary deposition. The sites had to have geological evidence of being relatively young, to maximize the chance of finding bones dating to near the final extinction; they had to span the full range of Australia's Late Pleistocene habitats, climate zones and geographical regions; and they had to reflect all major types of fossil site.

The results were clear: all of these putatively young sites containing articulated skeletons are at least 46,000 years old. Statistical analysis yielded a maximum-likelihood estimate of the extinction time itself as 46,000 years, with a 95% confidence interval of 51,000 to 40,000 years. Much younger dates (as recent as 2,000 years ago) were obtained for sediments containing disarticulated bones, but that is surely because old bones had been washed into young sediments.

Four conclusions follow. First, extinction occurred within a short time across the range of Australia's latitudes, longitudes, habitats, climate zones and megafaunal species. So the extinctions must have resulted from one common ultimate cause, not from agents peculiar to certain habitats, areas or species. More precise dates will be required to see whether extinction times differed slightly between Australia and New Guinea (which were joined at times of low sea level during the Late Pleistocene), or between Australia's northwest coast (the likely site of human arrival) and its southeast or interior.

Second, the extinctions occurred during a phase of benign moist climate, 25,000 years before Australia's arid phase at the Last Glacial Maximum. So that single ultimate cause is unlikely to have been climatic deterioration, as some authors have argued.

Third, the estimated time of extinction (46,000 ± 5,000 years ago) is only shortly after the estimated date of human arrival (56,000 ± 4,000 years ago). The statistical margin of errors means one cannot yet conclude whether the megafauna survived human arrival by just 400 years or by over 10,000 years. But we can surely implicate human arrival as the ultimate cause of the extinctions, though perhaps in diverse ways. For instance, overhunting of naive prey such as large herbivores would have resulted first in their extinction and then in that of large predators (losing their food source is a more likely explanation for the demise of 1-ton carnivorous lizards than their being clubbed out of existence by humans). Alternatively, extinction of the dominant herbivores, or fires started by the human immigrants, could have driven serious changes in habitat with the same end result.

What about other places — such as the Bismarcks, Cyprus, West Indies, Fiji, Madagascar and New Zealand — where megafauna similarly evolved in the absence of humans? The estimated dates of extinctions on these six islands fall in that sequence from 32,000 to just 700 years ago, and do not coincide with each other, or with Australia's extinction event, or even with the end of the last ice age around 10,000 years ago. But in each case they did occur soon after humans arrived in the area concerned, suggesting that these extinctions were also triggered by the arrival of humans.

In all of these cases, like that of Australia, human arrival and the associated megafaunal extinction did not occur at times of major climate change. But that is not so for the Americas, which lost an even richer megafauna than Australia. In North and South America, humans probably did us the disservice of arriving, and the megafauna of dropping dead, at the end of the last ice age. Some scientists, who have focused on the American extinctions alone and ignored the extinctions elsewhere in the world, are impressed only by the coincidence between the ends of the ice age and of the American megafauna, and not by the likely coincidence also with the arrival of innocent vegetarian humans. Other scientists see the Americas as fitting the worldwide pattern of coincident human arrival and megafaunal extinction — the only difference being that both events in the Americas (but not elsewhere) happened at a time of major climate change.

Like the finale of Mozart's Jupiter Symphony, debate about the American extinctions revolves around several themes. Here, the leading side theme is the uncertain arrival date of humans in the Americas. Some scientists especially invoke the much-debated ^{14}C dating of Chile's Monte Verde site (which, incidentally, needs to be re-dated by optical luminescence). They conclude that humans reached the Americas long before the ice age ended or the megafauna became extinct.

© 2001 Macmillan Magazines Ltd

Others, noting the hundreds of sites with Clovis-type artefacts scattered all over the continental United States and dating at the end of the ice age, conclude that Clovis sites really do mark the first settlement of the Americas south of Canada. If (say this second group of authors) there really had been humans in the Americas long before Clovis times, we would by now have discovered hundreds of unequivocal pre-Clovis sites, as we have in Australia and Europe. Palaeontologists and archaeologists who work on the prehistory of the Americas will no doubt continue to argue about these matters after reading the paper by Roberts *et al.*[2]. ■

Jared M. Diamond is in the Department of Physiology, University of California Medical School, Los Angeles, California 90095-1751, USA.
e-mail: jdiamond@mednet.ucla.edu

1. Martin, P. S. & Klein, R. G. (eds) *Quaternary Extinctions: A Prehistoric Revolution* (Univ. Arizona Press, Tucson, 1984).
2. Roberts, R. G. *et al. Science* **292,** 1888–1892 (2001).
3. Baynes, A. *Rec. W. Aust. Mus.* (suppl.) **57,** 391 (1999).
4. Roberts, R. G. *et al. Quat. Sci. Rev.* **13,** 575–583 (1994).
5. Miller, G. H. *et al. Science* **283,** 205–208 (1999).

Neurodegenerative diseases

Origins of instability

Richard R. Sinden

Errors in DNA replication are thought to underlie the lengthening of tracts of repeated DNA that occurs in some neurodegenerative diseases. But mechanisms for repairing damaged DNA may also be responsible.

Many neurological and neurodegenerative diseases, such as Huntington's disease and fragile X syndrome, share a similar genetic basis — the lengthening of tracts of repeated DNA sequence. The molecular mechanisms that bring about this 'repeat instability' have attracted much attention, but have remained somewhat mysterious. At a recent meeting*, however, our understanding of these mechanisms blossomed with the revelation that proteins that repair damaged DNA may be responsible for generating repeat instability in cells that are not dividing.

At least 14 human diseases have been associated with the lengthening (expansion) of tracts of nucleotide triplets, such as CTG (CAG on the complementary DNA strand), CGG (CCG) and GAA (TTC), in various human genes[1]. As well as Huntington's disease and fragile X syndrome, these diseases include myotonic dystrophy type 1, several spinocerebellar ataxias and Friedreich's ataxia. Breaking the 'triplet-repeat' mould, spinocerebellar ataxia type 10 is linked to the expansion of pentanucleotide ATTCT (AGAAT) repeats, from about 14 to 4,500 repeats[2]. The pathological bases of these diseases vary, but repeat expansion is the underlying mutation in all.

Two principal classes of repeat instability are associated with these diseases. First, when the tracts of repeats occur within non-protein-coding regions of an affected gene, the tracts can expand by a factor of 10–20, or even more, between generations. This large-scale expansion occurs in fragile X syndrome, myotonic dystrophy, spinocerebellar ataxias 8 and 10, and Friedreich's ataxia. Massive expansion of these particular repeats has been seen only in humans (although other repeats expand in mice[3]). The second class involves small expansions. Small-scale expansion of CAG repeats can cause disease when the repeats encode a tract of polyglutamine amino acids within an affected protein; for example, the presence of 30 CAG repeats in the gene encoding the huntingtin protein is in the normal range, whereas more than 36 CAG repeats cause Huntington's disease. In some disorders, such as myotonic dystrophy type 1 and Huntington's disease, short expansions accumulate differently in different cells and tissues throughout life. Small expansions may also occur between generations.

Conventional wisdom says that small expansions or deletions within tracts of repeats are generated during genome replication — that is, as cells divide. During DNA replication, a double-stranded DNA helix must first be unwound and unzipped; the two separated strands are then used as templates for the production of two new strands. When the DNA includes long tracts of simple repeats, it is easy for the growing DNA chain and the original template to become misaligned if DNA-replicating enzymes pause and separate from the template. This can lead to deletions or expansions of sequence when replication recommences. In fact, in a primate replication system, there is a marked bias for expansion from 79 to 107 repeats of the CTG repeat that is found in myotonic dystrophy (C. Pearson, Hospital for Sick Children, Toronto). Moreover, disease-associated triplet repeats can form 'alternative' DNA conformations, including hairpins and slipped-strand structures, as well as intramolecular triplex and quadruplex DNA structures. These structures might promote fractious replication.

It seems, however, that theories involving genome replication cannot account for all small-scale repeat instability. In mouse models of neurodegenerative disease, instability generally occurs at different rates in different tissues, with continued expansion occurring as the mice age. Instabilities can also show a dependence on their chromosomal context. Most remarkably, it seems that repeat variations can occur in the absence of the DNA replication that is associated with genome duplication. In mice engineered to provide models of myotonic dystrophy or Huntington's disease, instability often occurs to a greater degree in cells that do not divide (such as brain and kidney cells) compared with cells that divide frequently[4–6]. Variations in instability are also observed in cultured cells[7]. Moreover, repeat variability has also been seen in eggs[8] and developing sperm[9], which do not replicate and do not undergo a process called DNA recombination. Both replication

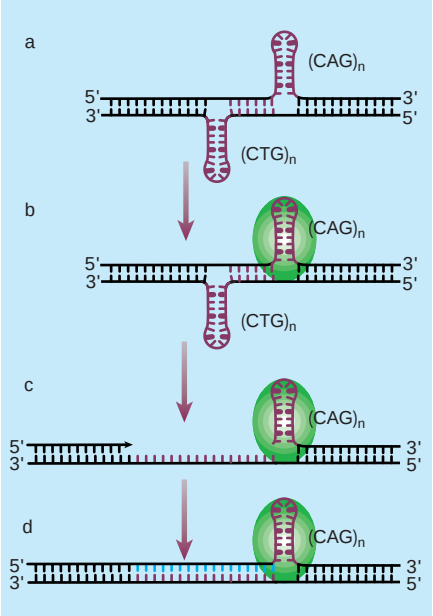


Figure 1 **In one model for the production of unstable tracts of nucleotide repeats in cells that are not dividing, 'alternative' DNA structures are crucial.** a, **During DNA replication, the individual DNA strands can become separated and then paired up again. During this process, slipped-strand DNA structures, as shown, or other 'alternative' structures may form; this is particularly likely when tracts of nucleotide repeats (such as CTG, or CAG on the complementary strand; purple) are present. These structures may also form spontaneously.** b, **The DNA-repair proteins MSH2 or MSH3 (green oval) bind to and stabilize the hairpin in one strand (here the upper strand).** c, **The repair proteins introduce an incision next to the upper hairpin, allowing the strands to separate and the lower hairpin to unfold.** d, **The resulting gap in the upper strand is filled in (turquoise) by using the lower strand as a template. The tract of repeats is thereby lengthened relative to the original DNA. (The thicker tick marks denote T–T and A–A mismatches that occur in the hairpin stems.)**

**Third International Conference on Unstable Microsatellites & Human Disease*, Noordwijerhout, The Netherlands, 21–25 April 2001.

© 2001 Macmillan Magazines Ltd

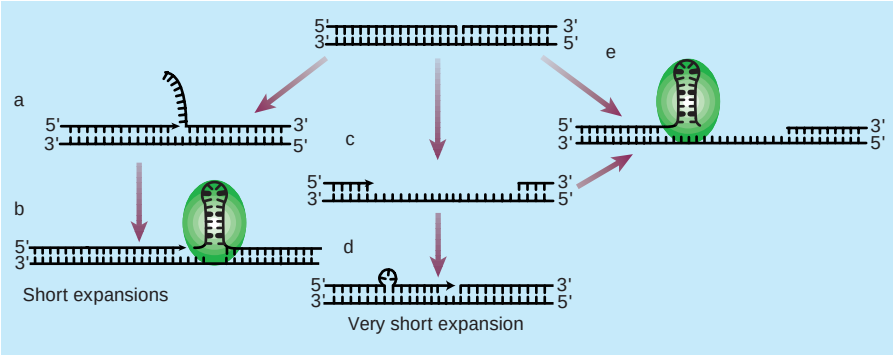


Figure 2 **A nick in DNA could promote the instability of tracts of nucleotide repeats.** a, **During the repair of the gap, too much of the lower strand might be copied, producing a flap. This is 'strand-displacement synthesis'.** b, **Continued copying will lead to the formation of a hairpin within the flap. The repair proteins MSH2 or MSH3 might bind to and stabilize the hairpin; sealing the nick will then lead to an expansion in the upper strand. The next time the cell duplicates its DNA, the expansion will be copied if unrepaired, and both strands will contain the expansion.** c, **A large gap might form as a result of the initial nick.** d, **The error-prone replication enzyme DNA polymerase-β (or another polymerase) fills in the gap; if it slips, it could cause addition or deletion of one or more repeats.** e, **DNA 'melting' at the nick could lead to hairpin formation, perhaps stabilized by MSH2/MSH3, creating a gap. DNA melting and hairpin formation could also occur at a gap, as well as a nick. The hairpin may form at either end of the gap. If the gap is filled without removing the hairpin, repeat expansion will occur.**

and recombination are thereby excluded as obligatory pathways to repeat variability.

What mechanism, then, can explain the age-dependent accumulation of mutations in non-dividing cells? A key revelation was that mice with Huntington's disease that lack the DNA-repair protein MSH2 do not show age-dependent repeat variability in different tissues[10]. Studies of animal models of myotonic dystrophy that lack the DNA-repair proteins MSH2 (G. Gourdon, Hôpital Necker-Enfants Malades, Paris), PMS2 (D. Monckton, Univ. Glasgow) or MSH3 (B. Wieringa, Univ. Nijmegen), as well as further studies of mice with Huntington's disease[9], have now confirmed that age-dependent repeat instability requires repair proteins. Moreover, MSH2 is required for repeat instability in developing sperm in mice with Huntington's disease[9]. However, instability still occurs in animals that lack the repair protein MSH6 (B. Wieringa) or the recombination proteins RAD52 and RAD54 (G. Gourdon).

There are several ways in which DNA can be altered that might lead eventually to DNA-repair-dependent repeat instability. First, alternative DNA structures, such as slipped-strand DNA, might form within repeat tracts after DNA replication. These structures could then bind to MSH2 or MSH3, which would try to mend the unusual structures, perhaps (in fairly complicated ways) resulting in errors such as repeat expansion (Fig. 1; C. Pearson). Second, DNA sustains continuous spontaneous damage, for example as a result of oxidation[8]. Repeat tracts or alternative structures might be particularly sensitive to damage — and so particularly dependent on effective DNA repair.

Third, single- or double-stranded breaks can occur in DNA, either *de novo* or as a result of incomplete DNA replication. Moreover, larger gaps could accumulate, with flaps, hairpins or other complex structures forming at their ends (C. McMurray, Mayo Foundation, Rochester, Minnesota). DNA-repair proteins might bind to and stabilize these unusual repeat structures, indirectly promoting expansion (Fig. 2). In the case of mice with Huntington's disease, the repeat tracts expand by just one to three repeats[9]; this may be too few to allow a stable repeat secondary structure to form. Here, instability might instead reflect the slippage of replicating enzymes during an error-prone repair process.

With this new knowledge in hand, it may one day prove possible to develop approaches that minimize the lengthening of repeat tracts that can occur throughout life. But another important part of the story — the molecular basis of the large-scale tract expansion that can occur between generations — remains a mystery. ■

Richard R. Sinden is at the Center for Genome Research, Institute of Biosciences and Technology, Texas A&M University System Health Sciences Center, 2121 West Holcombe Boulevard, Houston, Texas 77030-3303, USA.
e-mail: rsinden@ibt.tamu.edu

1. Cummings, C. J. & Zoghbi, H. Y. *Hum. Mol. Genet.* **9,** 909–916 (2000).
2. Matsuura,T. *et al. Nature Genet.* **26,** 191–194 (2000).
3. Bois, P. R., Southgate, L. & Jeffreys, A. J. *Mamm. Genome* **12,** 104–111 (2001).
4. Kennedy, L. & Shelbourne, P. F. *Hum. Mol. Genet.* **9,** 2539–2544 (2000).
5. Seznec, H. *et al. Hum. Mol. Genet.* **9,** 1185–1194 (2000).
6. Fortune, M. T., Vassilopoulos, C., Coolbaugh, M. I., Siciliano, M. J. & Monckton, D. G. *Hum. Mol. Genet.* **9,** 439–445 (2000).
7. Gomes-Pereira, M., Fortune, M. T. & Monckton, D. G. *Hum. Mol. Genet.* **10,** 845–854 (2001).
8. Kaytor, M. D., Burright, E. N., Duvick, L. A., Zoghbi, H. Y. & Orr, H. T. *Hum. Mol. Genet.* **6,** 2135–2139 (1997).
9. Kovtun, I. V. & McMurray, C. T. *Nature Genet.* **27,** 407–411 (2001).
10. Manley, K., Shirley, T. L., Flaherty, L. & Messer, A. *Nature Genet.* **23,** 471–473 (1999).

Daedalus

The tense road

A road, says Daedalus, is quite a clever mechanical invention. It takes the force applied by a driven wheel, and reacts to it by momentary deformation. The resulting reaction of the road on the vehicle drives it along. Fast or heavy vehicles, by the powerful reaction they impose on the road, soon damage it. Their powered wheels do far more damage than the merely passive load of the coasting wheels. Daedalus sees the force of the engine being applied first to the tyre, and only then to the road. The tyre is canted, at least at the bottom, and the angle of cant applies a force to the vehicle.

So Daedalus wonders what would happen if the tyre were made from his 'slow rubber' of last week. At a slow enough speed, nothing would happen, and the vehicle would proceed with normal engine efficiency along the road. But at some well-defined speed, the slow restitution of the rubber would hold the canting effect well beyond the time of contact. The tyre would become very lossy, and most of the car's energy would appear not in acceleration, but as heat in the tyres. (Carbon black is put into tyres specifically to dissipate heat.)

Thus a tyre of a properly compounded slow rubber would act as an automatic speed limit. On the principle 'never do by law what you can do by engineering', Daedalus advocates the sale of 'slow tyres' to prove to the police that a specific car could never have exceeded the speed limit.

But clearly the law would be better served by a 'slow road'with the equivalent property. Any car or lorry, no matter what its tyres, would then feel a dramatic slowing on passing onto such a road surface. The road would be deformed by the driven tyres, and this deformation would be restored too late to power them along. Instead, the road would get hot.

The slow road might need to be replaced at frequent intervals, as it absorbed the power of many roadhogs, but it could still be a sound investment. On roundabouts, junctions and sharp bends, and on motorways (with the exception of overtaking lanes), it would restrict motorists to the speed limit, without requiring any legal action. A driver would notice nothing, except that no amount of throttle could accelerate him above the legal limit on such stretches of road. Daedalus hopes that such stretches will hardly get warm from the few roadhogs who challenge them, and that their deterioration (thermal and mechanical) will be very slight.

David Jones

© 2001 Macmillan Magazines Ltd

The ins and outs of signalling

Julian Downward

The study of how cells communicate impinges on all aspects of biology, from development to disease. At first glance it's a horrendously complicated business, but some simple themes are emerging.

All cells must continually sense their surrounding environment and make decisions on the basis of that information. Single-celled organisms must be able to tell which nutrients are nearby and regulate their metabolic processes accordingly. Cells in multicellular organisms such as ourselves must sense the presence of neighbouring cells and hormones when making decisions such as to whether to proliferate, move or die. These processes all require the transfer of information (Fig. 1) from detection systems referred to as receptors through intermediate molecules within the cell, to cause changes in the expression of genes and the activity of enzymes — the specialized protein machinery that carries out all of a cell's functions. 'Signal transduction', 'cell signalling' or simply 'signalling' is the study of the mechanisms by which this transfer of biological information comes about.

Signalling can be studied at the level of the individual cell or the whole organism. For individual cells, signalling is crucial to decisions about division, specialization, death and metabolic control. In more specialized cells it is central to immunity and the transmission of nerve impulses, to name but two examples. At the level of whole multicellular organisms, signalling controls growth and development, as well as aspects of metabolism and behaviour. Not surprisingly, then, signalling malfunctions underlie many human diseases (Box 1, overleaf).

The history of the study of biological signalling can be read in the Nobel Prizes for Physiology or Medicine awarded over the past 30 years, starting with Sutherland in 1971 and culminating with the 2000 prize to Carlsson, Greengard and Kandel for their studies of signal transduction in the nervous system. The work of these and many other scientists has established the outlines of numerous signalling pathways. These vary enormously in their details. Yet despite this diversity, a much simpler underlying logic is beginning to emerge. Here I aim to explore key themes in biological signalling, and to illustrate them with some of the better-understood signalling systems. I hope that the reader will be convinced that signal transduction is not the impenetrable soup of acronyms that it might at first appear to be.

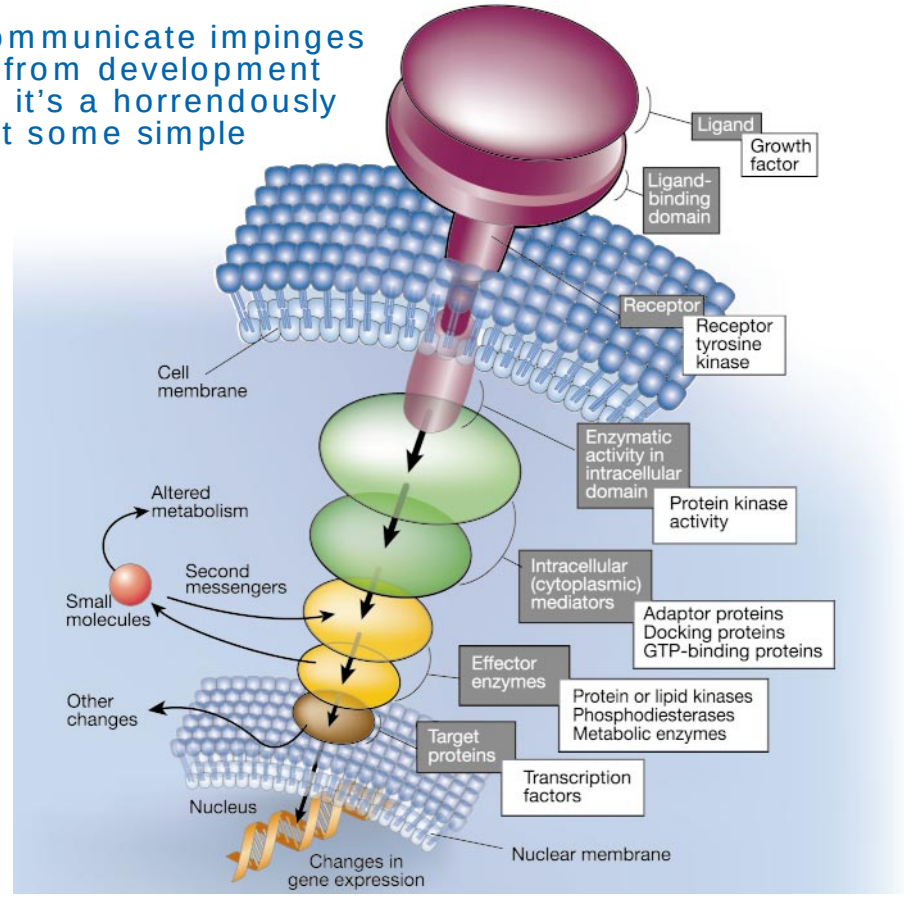


Figure 1 **A generic signalling pathway. The grey boxes indicate general components of signalling pathways; the white boxes show specific examples. On the outside of the cell membrane, receptors bind to biologically active ligands such as growth factors. As a result, enzymatic activity associated with the intracellular part of the receptor is altered. This can affect the association of the receptor with intracellular mediators, or the localization or function of those mediators. These in turn alter the activity of 'effector' enzymes. Some effectors can move to the nucleus and control gene expression, or they can induce other proteins to do so. Others target small molecules, either generating further signalling mediators (second messengers) or controlling the metabolic state of the cell. Real signalling pathways may bypass entire classes of these molecules, or may have several components in one or more class, working either in series or in parallel.**

Signalling nuts and bolts

Receptors. Signalling pathways start with receptor proteins that are able to sense a change in the environment outside the cell. Receptors are most commonly found at the cell surface, where they bind to extracellular molecules that cannot penetrate the plasma membrane — the lipid boundary between the cell and the outside world. Receptors on the cell surface can bind to water-soluble signalling proteins such as growth factors and peptide hormones, which may be produced at a distant site in the body (and delivered in the bloodstream) or by neighbouring cells. Receptors can also bind to small water-soluble molecules such as nutrients.

As a result of engagement of these receptor proteins by their binding partners ('ligands'), a signal is transferred across the plasma membrane[1]. In most cases, the receptor itself spans the membrane. Ligand binding causes a change in the shape of the protein; this change is transmitted from the extracellular part of the receptor to the part inside the cell. Sometimes this involves the formation of dimers of receptor molecules — two receptors bonded together. This tends to occur for receptor proteins with amino-acid chains that cross the membrane only once. For receptors that fold up so that they span the membrane several times, ligand binding may cause different parts of the molecule to reorientate themselves with respect to each other.

Inside the cell, the change in the recep-

© 2001 Macmillan Magazines Ltd

tor's conformation can result in the stimulation of an activity that is integral to the intracellular part of the receptor. One of the most common, and perhaps the best studied, activities is that of the kinases — enzymes that catalyse the transfer of a phosphate group from adenosine triphosphate (ATP, the cell's soluble energy-storing molecule) to protein substrates. Two principal classes of kinases exist: those, such as members of the activin/transforming growth factor (TGF)-β receptor family, that transfer phosphate to serine and threonine amino-acid residues in target proteins; and those, including the growth-factor-receptor tyrosine kinase family, that target tyrosine residues[2,3].

Other receptors have different activities. For example, some act as ion channels and pumps, allowing ions into or out of the cell to control cellular functions. Alternatively, the part of the receptor within the cell may interact with other signalling proteins and pass on the information to them by inducing changes in conformation and activity. A common example is provided by the abundant family of 'G-protein-coupled' receptors.

Cell-surface receptors can bind to insoluble structures as well as to soluble molecules. For example, integrins are receptors that bind to the extracellular matrix[4] — the glue that holds the body together. Cell-adhesion molecules are responsible for binding cells to their neighbours[5]. These types of receptor are important not only for structural reasons — in establishing tissue architecture — but also in informing cells of the presence of other cells and of the matrix, and enabling them to behave accordingly, for example by ceasing to grow when they are surrounded by other cells. Finally, receptors inside the cell bind to signalling molecules that can penetrate the plasma membrane, such as steroid hormones and nitric oxide.

Intracellular mediators. Inside the cell, receptors pass on the signal to other molecules. As mentioned above, receptor kinases catalyse the transfer of a phosphate group onto substrate proteins, which may be enzymes with activities that are directly affected by this phosphorylation. For example, receptors for TGF-β phosphorylate proteins of the SMAD family (which are transcriptional regulators that work to control gene expression), resulting in their activation[6]. Receptors themselves are often modified by self-phosphorylation; this occurs for growth-factor receptors that stimulate phosphorylation on tyrosine residues within themselves.

The phosphorylation of substrate proteins can also affect their interactions with other molecules. For example, phosphorylated residues in a protein can act as binding sites for specific recognition domains in other proteins[7]. A domain in a protein is a self-folding unit with a particular sequence and conformation, and certain domains allow proteins to recognize each other. So, as a result of phosphorylation, for example, protein complexes can assemble, resulting in changes in the localization or activity of enzymes.

Some proteins in these complexes, referred to as 'adaptor' or 'docking' proteins, may work only to bring together other signalling molecules. One example is provided by the Ras pathway (Fig. 2). This pathway has been particularly well studied, not least because mutation of the *ras* gene occurs in many cancers. In this pathway, binding of a growth factor to its cell-surface receptor leads to the self-phosphorylation of the receptor. The SH2 domain of an adaptor protein, Grb2, grabs on to the phosphorylated tyrosine residue; Grb2 thereby links the activated receptor to an enzyme, Sos, that regulates Ras. As both the receptor and Ras are found at the plasma membrane, the outcome of the assembly of this complex is that Sos has increased access to Ras[8]. Ras then activates a cascade of intracellular protein kinases, culminating in activation of a kinase called ERK (a so-called MAP kinase), which directly influences gene expression.

Nuclear events. Like the Ras pathway, many signalling pathways end with a change in the gene-expression programme of the cell[9]. This usually requires the movement of a protein from the body of the cell to the nucleus in response to activation of the signalling pathway. Specific recognition systems control the import and export of proteins to and from the nucleus. These systems recognize sequence motifs in the proteins, and the accessibility of the motifs may be altered as a result of phosphorylation or complex formation[10]. In the growth-factor-activated Ras pathway, ERK moves to the nucleus as a result of its phosphorylation and activation by the previous kinase in the cascade. Once inside the nucleus, ERK phosphorylates and activates proteins that are involved in the transcription of genes into messenger RNAs. These mRNAs may then be translated into protein, so altering the protein composition of the cell and leading to changes in cell function.

Signalling concepts

Specificity. The route that information takes from outside a cell to the endpoint of a signalling pathway is defined by the molecular interactions of the proteins in the pathway. These interactions are therefore critical. They are frequently induced only upon activation of the pathway, and they must be sufficiently

Box 1 Signalling and disease

Many diseases involve malfunction of signalling pathways. In particular, much of the basic work on the regulation of cell proliferation has been carried out to obtain a better understanding of cancer. Cancer is a disease of signal malfunctioning that is driven by microevolutionary processes. Once a cell suffers an inactivating mutation in a growth-inhibitory pathway (that is, in a tumour-suppressor gene), or an activating mutation in a growth-promoting pathway (in other words, in an oncogene), it will have a competitive proliferative advantage over its neighbouring cells[23]. This can lead eventually to malignant growth of a mutant clone of cells at the expense of the whole organism. To avoid cancer, animals have developed several signalling pathways that prevent uncontrolled cell proliferation, meaning that only after several mutations have occurred in the same cell does cancer result.

The study of oncogenes has led to the identification of many signalling proteins that are involved in cell growth, including growth factors, their receptors, intracellular mediators and transcription factors. Likewise, the study of tumour-suppressor genes has led to the identification of proteins that negatively regulate cell-cycle progression, cell survival and the ability to invade surrounding tissue.

Many other non-infectious diseases are caused by defects in signalling pathways. Diabetes, for example, results from defects in the insulin-signalling pathway used to control blood glucose levels[24]. Insulin works through a receptor-tyrosine-kinase signalling pathway that results in activation of phosphatidylinositol-3-OH kinase, Akt and other intermediate enzymes, particularly in fat and muscle cells. Ultimately, these control the amounts of the glucose receptor at the cell surface, as well as the levels of various metabolic enzymes.

Moreover, many developmental diseases result from signalling defects, one example being forms of achondroplasia (dwarfism) in which the receptor tyrosine kinase for fibroblast growth factor is mutated. Immunological disorders can also have a basis in signalling malfunction. For example, in agammaglobulinaemia (which is characterized by a lack of immunoglobulin antibodies in the blood), mutation in the B-cell tyrosine kinase Btk often occurs, resulting in a failure of this enzyme to respond to activation of phosphatidylinositol-3-OH kinase.

We can hope that in-depth characterization of signalling pathways will lead eventually to an ability to intervene in diseases in which those pathways are defective. For example, a drug called STI571, which inhibits the tyrosine kinase Abl, is now being used successfully to treat certain leukaemias in which this enzyme has been erroneously activated[25]. **J.D.**

© 2001 Macmillan Magazines Ltd

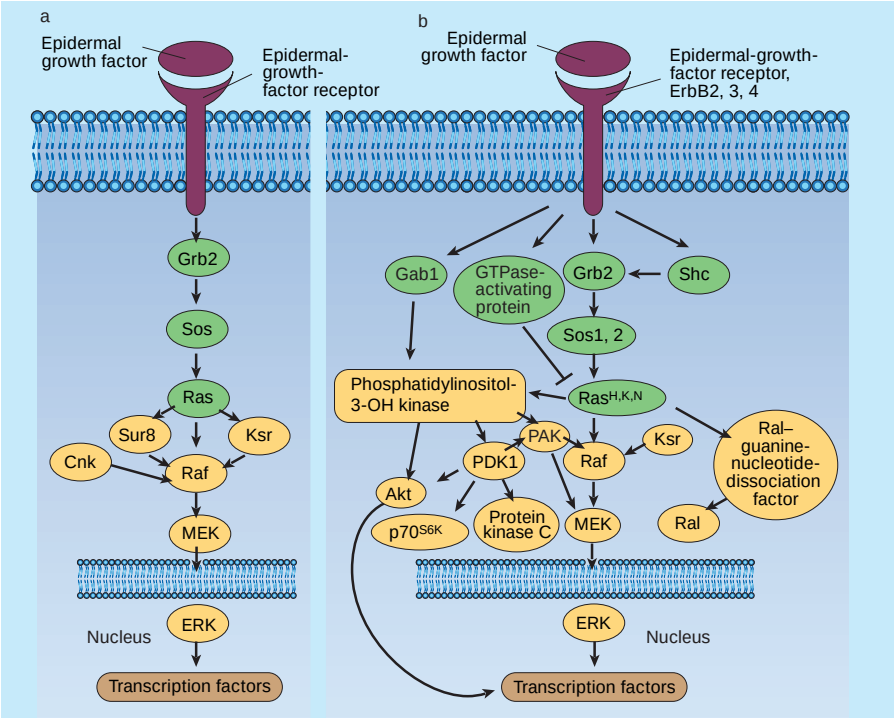


Figure 2 **Two ways of looking at one signalling pathway — the Ras pathway — from growth factor to changes in gene expression.** a, **The genetic view is based largely on studies of the development of flies, worms and, to some extent, mice.** b, **The molecular cell-biological view is derived mainly from work on mammalian cells in tissue culture. The genetic approach gives a simpler picture, but principally considers only components that are essential for normal development of the organism under laboratory conditions. Being based largely on simple model organisms, genetics has been less useful in studying the function of the several closely related members of gene families that are typically found in mammals (such as the epidermal-growth-factor receptor and Erb2, 3 and 4). The molecular cell-biological view is more complex: it is pieced together from experiments done under different conditions and in different cell types, often relying on overexpression of proteins in the pathway. It reveals many possible subtleties of the pathway, but these are unlikely to occur together to a significant extent at the same time in the same cell. The most effective modelling of signalling pathways draws on experimental evidence of both types to reach a mutually consistent picture.**

specific to ensure that signals do not become misrouted. How is specificity achieved?

A strong theme that has emerged over the past decade is that proteins are made up of modular units, which belong to families of related structures[11]. Modules may be either domains or smaller motifs, such as some phosphorylation sites. Modules can direct protein–protein interactions through their ability to interact with other modules (Fig. 3, overleaf). They may also have enzymatic activity, which in signalling proteins is often used to regulate other molecules. A large protein may contain several different modules, each of which behaves similarly to related modules in other proteins. A protein is thus made up of generic functional building blocks, which have been shuffled around during evolution to yield different combinations of interactions and activities.

To ensure the accuracy of information flow, irrelevant interactions between proteins need to be kept at a manageable level. The use of interacting modules narrows down the number of connections that can occur. But there is only a limited number of module families, and many proteins will contain the same modules, so how can they distinguish between related targets? Several of the design features of signalling pathways may aid specificity, including the use of combinations of modules to achieve higher selectivity for a given protein–protein interaction, and of scaffolds of adaptor proteins to help to bring partners together[11,12]. The localization of proteins to specific cellular compartments can also limit the range of potential interactions.

Signalling pathways presumably evolved to transfer information with acceptable selectivity under normal conditions. But a common experimental technique for analysing signalling pathways involves overexpressing either normal or mutant proteins to see the effects on the cell. Unfortunately, as many proteins contain the same modules, this technique may well result in perturbation of pathways other than the one that contains the protein of interest. It is not clear how big a problem this is. But there is a real danger that many data from overexpression experiments will have to be re-evaluated if we are to obtain a clear overall picture of cell signalling.

Complexity. The size of even the simplest genome of a unicellular organism allows for enormous complexity of signalling. A sizeable proportion of the genome of a multicellular organism encodes signalling molecules; for instance, over 2% of the genome of the worm *Caenorhabditis elegans* encodes its 400 kinases[13]. Evolution drives the generation of robust, adaptable systems, and the modular nature of signalling proteins makes it easy to generate new variants to meet new environmental challenges. There is probably little benefit to be had from simplicity in itself, as the energy required to make so many genes and their products is amply repaid by the improved performance of the organism. Often, organisms have developed several ways to do the same thing, and this redundancy perhaps reflects the importance of some signalling pathways. Nonetheless, it is likely that the apparent complexity of many published signalling diagrams may be at least partly due to the experimental techniques used (Fig. 2).

Signal integration. Cells receive inputs from many signalling pathways at the same time and must interpret them together, in the context of each other, before making decisions. There are several known ways in which cells do this, although this is an area where much work remains to be done.

For example, individual enzymes can receive input from several pathways, which often feed into different modules of the same enzyme. This type of integration process is known as crosstalk, whereby an important signalling pathway is influenced by the activity of another, even if the second pathway cannot fully control the first. Examples of this include the ability of phosphatidylinositol-3-OH kinase and another kinase, PAK, to feed into the MAP kinase pathway that is triggered by activated Ras.

As well as crosstalk between pathways, feedback loops can occur, in which a component further on in a pathway either positively or negatively influences the activity of an earlier component. Negative-feedback systems are common in signalling. In contrast, positive-feedback systems are rarer, as they can be inherently unstable. However, it is common to find a degree of amplification built into signalling pathways, such that each step in the pathway results in an increase in signal strength, turning a weak signal at the cell surface into a strong one in the nucleus.

Signal integration also occurs at the level of gene expression. Different signals lead to activation of different sets of transcription factors, so affecting the expression of different genes. The cell's overall response will depend on how the induced proteins interact with and influence each other.

Localization and translocation. The location of a signalling protein within a cell affects the molecules with which that protein can interact. Movement of a signalling protein to a different cellular location — translocation — may result in a change in its interacting part-

© 2001 Macmillan Magazines Ltd

ners and hence in its activity. Regulation of the localization of signalling proteins is a key aspect of many signal-transduction pathways. Thus the activation of phosphatidylinositol-3-OH kinase by receptor tyrosine kinases leads to the generation of a specific lipid, phosphatidylinositol-3,4,5-trisphosphate, at the plasma membrane. The lipid binds to so-called pleckstrin-homology domains in proteins such as Akt, a serine kinase[14,15]. This causes Akt to translocate to the plasma membrane, where it comes into contact with another protein kinase that phosphorylates and activates it. Phosphorylated Akt can then move to other sites in the cell, including the nucleus, while maintaining its activity.

Control of localization is also important in the control of many transcription factors, which can be kept out of the nucleus, and hence away from their target genes, by regulated interactions with other proteins. For example, the Forkhead transcription factors are inactivated when phosphorylated by Akt. Phosphorylation creates a binding site for 14-3-3 proteins, impairing the import of these transcription factors into the nucleus[14].

Signal amplitude and duration. The response of cells to activation of a particular signalling pathway depends on the strength of that activation. Pathways can show graded responses, like a rheostat — the stronger the activation of the intermediate proteins in the pathway, the stronger the final activity. In these cases, because different cells may show different sensitivities to a signal, low signal strengths might activate a subset of the responses that are activated by high signal strengths. In addition, some pathways work as on/off switches — once signal strength rises above a certain level, positive feedback results in full activation of downstream targets.

The time course of a signalling pathway can also be critical. Transient activation of a pathway may have quite different effects to long-term activation. For example, in cultured PC12 cells, epidermal growth factor induces transient activation of the MAP kinase pathway, causing the cells to multiply slowly. Nerve growth factor, in contrast, induces sustained activation of the MAP kinase cascade, and the cells become specialized to resemble nerve cells[16]. During the stimulation of cell proliferation by growth factors, signals may be needed at specific periods of the cell-division cycle but not at others. What's more, signals may be interpreted differently at different points in the cell cycle, because of changes in other combinations of signals.

Partner 1	Partner 2	Interaction used by
Phosphorylated tyrosine residues	SH2 domain PTB domain	Growth-factor receptor tyrosine kinase
Phosphorylated serine/threonine residues	FHA domain 14-3-3 domain	Enzymes within signalling pathways, such as Raf Transcription factors
Motifs rich in proline residues	SH3 domain WW domain	Adaptor proteins regulated by growth factors Also used in regulating cytoskeletal, cytoplasmic and nuclear proteins
Phosphoinositide lipids	PH domains	Growth-factor-regulated enzymes, such as Akt
DD, DED, CARD domains	DD, DED, CARD domain	Used in regulation of programmed cell death
G(S/T)XVI sequence (at carboxy terminus)	PDZ domain	Ion channels Also used for control of cell polarity

Figure 3 **Some common protein-interaction modules used in signalling. Modules can direct protein–protein interactions through their ability to specifically recognize other modules. Abbreviations: SH2 and SH3, Src-homology domains 2 and 3; PTB, phosphotyrosine-binding domain; FHA, Forkhead-associated domain; WW, tryptophan–tryptophan domain; PH, pleckstrin-homology domain; DD, death domain; DED, death-effector domain; CARD, caspase-recruitment domain; G(S/T)XVI, amino-acid sequence glutamate, serine/threonine, any amino acid, valine, isoleucine; PDZ, a domain present in so-called PSD-95, Dlg and ZO1/2 proteins[7].**

The future

Our ability to understand signalling pathways will be revolutionized by the availability of full genome sequences. Complete annotated genome sequences are available for a yeast, a nematode worm, a mustard plant and a fruit-fly, and the sequencing, assembly and annotation of the human genome are approaching completion[17,18]. These innovations have fundamentally redefined the study of all aspects of biology, including signalling. In yeast, worms and flies, we can now find every member of every family of signalling proteins, and this is rapidly becoming the case for humans as well. As more is learnt, it will be possible to fill in more and more detail about the function of as-yet-uncharacterized proteins, simply from the sequence of their genes.

Of course, knowledge of gene sequences will not in itself fully define protein function. But such knowledge will allow the creation of comprehensive genome-wide databases of signalling information. For example, protein-interaction maps are being drawn for all the signalling components encoded by yeast and worm genomes; this approach will generate a matrix of all possible protein interactions in these organisms[19]. And the function of each gene in these organisms is being systematically knocked out, one at a time, in the effort to learn more[20]. Another approach is to use DNA microarrays to study the level of gene expression in response to activation of different pathways or deletion of different genes[21]. For example, by searching for similarities in gene-expression profiles between mutant yeast lacking a specific gene and normal yeast treated with a drug of unknown function, it can be possible to determine how and in what pathway the drug works.

Ultimately, genomic approaches like this, combined with conventional techniques, may lead to the generation of comprehensive diagrams of cellular signalling pathways. But we will have to learn a great deal more about handling complexity before we will be able to make complete sense of such a mass of data. It is likely that information in these quantities will never be easily understandable without the assistance of computer modelling. At present, modelling of signalling networks is in its infancy, although relatively simple parts of signal-transduction pathways have been modelled successfully[22]. Comprehensive web-based databases of information about signalling pathways are being built, and may in the future progress from being mere descriptive tools to predictive ones. ■

Julian Downward is in the Signal Transduction Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, London WC2A 3PX, UK.
e-mail: downward@icrf.icnet.uk

1. Schlessinger, J. *Cell* **103,** 211–225 (2000).
2. Hunter, T. *Phil. Trans. R. Soc. Lond. B.* **353,** 583–605 (1998).
3. Cohen, P. *Trends Biochem. Sci.* **25,** 596–601 (2000).
4. Schlaepfer, D. D. & Hunter, T. *Trends Cell Biol.* **8,** 151–157 (1998).
5. Daniel, J. M. & Reynolds, A. B. *Bioessays* **19,** 883–891 (1997).
6. Massague, J. & Chen, Y. G. *Genes Dev.* **14,** 627–644 (2000).
7. Pawson, T. & Scott, J. D. *Science* **278,** 2075–2080 (1997).
8. Shields, J. M. *et al. Trends Cell Biol.* **10,** 147–154 (2000).
9. Kornberg, R. D. *Trends Cell Biol.* **9,** M46–M49 (1999).
10. Christophe, D. *et al. Cell Signal.* **12,** 337–341 (2000).
11. Pawson, T. *Nature* **373,** 573–580 (1995).
12. Whitmarsh, A. J. & Davis, R. J. *Trends Biochem. Sci.* **23,** 481–485 (1998).
13. Plowman, G. D. *et al. Proc. Natl Acad. Sci. USA* **96,** 13603–13610 (1999).
14. Datta, S. R. *et al. Genes Dev.* **13,** 2905–2927 (1999).
15. Fruman, D. A. *et al. Annu. Rev. Biochem.* **67,** 481–507 (1998).
16. Marshall, C. J. *Cell* **80,** 179–185 (1995).
17. International Human Genome Sequencing Consortium *Nature* **409,** 860–921 (2001).
18. Venter, J. C. *et al. Science* **291,** 1304–1351 (2001).
19. Walhout, A. J. M. & Vidal, M. *Nature Rev. Mol. Cell Biol.* **2,** 55–62 (2001).
20. Oliver, S. G. *et al. Trends Biotechnol.* **16,** 373–378 (1998).
21. Schulze, A. & Downward, J. *J. Cell Sci.* **113,** 4151–4156 (2000).
22. Endy, D. & Brent, R. *Nature* **409,** 391–395 (2001).
23. Hanahan, D. & Weinberg, R. A. *Cell* **100,** 57–70 (2000).
24. Withers, D. J. & White, M. *Endocrinology* **141,** 1917–1921 (2000).
25. Mauro, M. J. & Druker, B. J. *Curr. Opin. Oncol.* **13,** 3–7 (2001).

Related websites

- **genome-www.stanford.edu/saccharomyces**
- **www.wormbase.org**
- **www.flybase.org**
- **www.ncbi.nlm.nih.gov/genome/guide/human**

© 2001 Macmillan Magazines Ltd

Knot tied around an octahedral metal centre

A zinc ion coordinates the folding of a linear polymer programmed to form a knot.

The topology of a molecule has a profound influence on its properties, so a range of sophisticated enzymes has evolved to manipulate molecular entanglements such as knots and chain links. Controlling the topology of synthetic oligomers remains a major challenge, and here we exploit the weak non-covalent interactions in such a molecule to tie a knot around a zinc ion. The folding of the knot is programmed into the covalent structure of the linear molecule, opening new routes for synthesizing macromolecular structures of unprecedented architecture and with well-defined topological properties.

The tools of molecular biology have been used to construct a range of closed knots from DNA and RNA[1,2], and optical tweezers have been applied to tie single DNA or protein molecules manually into open knots[3]. Synthetic strategies for the preparation of small-molecule knots have relied on the templating effects of non-covalent interactions to direct covalent-bond formation[4–7]. Attempts to exploit the ligand geometry around an octahedral metal centre to create a molecular trefoil knot[8] in conventional template synthesis have not been successful. We tested an alternative approach in which a metal ion is added to a presynthesized oligomer equipped with appropriate recognition sites that should cause spontaneous folding into an open knot[9].

We used molecular modelling to design two appropriate building blocks (Fig. 1a), a bidentate ligand (**1**) and a complementary rigid linker (**2**), which we covalently connected to give the linear oligomer **3**. Addition of one equivalent of zinc perchlorate, $Zn(ClO_4)_2$, to a solution of oligomer **3** in deuterodichloromethane caused spontaneous and quantitative conversion to a new species, which was accompanied by marked changes in the proton nuclear magnetic resonance (NMR) spectrum (Fig. 1b). Large shifts were evident in the aromatic signals and the methylene protons became non-equivalent, which suggests that the flexible chains are locked in a single conformation in the complex.

Fast-atom-bombardment mass spectrometry confirmed that a 1:1 complex had been formed and that no higher-order species were present. An X-ray crystal structure of the Zn**3**$(PF_6)_2$ complex (CCDC access no. 163073) proved that the open-knot architecture had been formed in solution. The non-covalent interactions that control folding are apparent in the structure shown in Fig. 1c: metal-ion coordination organizes the core of the knot, and the linker bipyridine subunits (**2**) straddle the ligands (**1**) as a result of aromatic interactions that organize the outside of the knot[10].

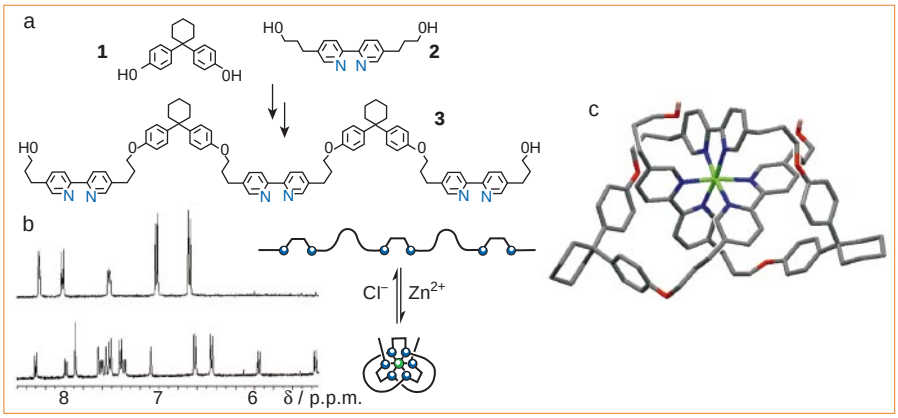


Figure 1 A synthetic oligomer that folds into a knot in the presence of zinc ions. **a,** Building blocks **1** and **2** were used to assemble oligomer **3**. **b,** Reversible folding of the knot complex. Part of the 400-MHz ^{1}H NMR spectrum of **3** in deuterodichloromethane (top) and the corresponding spectrum after addition of one equivalent of $Zn(ClO_4)_2$ (bottom; zinc ion represented in green) are shown. Addition of tetraethylammonium chloride to the zinc complex recovers the spectrum of the unfolded oligomer (top). **c,** Three-dimensional structure of the knot complex, Zn**3**$^{2+}$, as determined by X-ray crystallography (PF_6 anions and CH_3CN solvent molecules are omitted for clarity).

The folding process is fully reversible. Addition of zinc ions to oligomer **3** quantitatively folds the knot, addition of chloride quantitatively unfolds it to yield the free oligomer, and addition of silver ions (which precipitate the chloride) refolds it. The system can be repeatedly cycled through this process. The open-knot Zn**3**$^{2+}$ represents an important new synthetic intermediate for the assembly of a range of topologically complex molecular architectures. It can be readily prepared on a gram scale, and the terminal hydroxyl groups provide sites for elaboration. Macrocyclization would yield a closed trefoil knot, bulky stopper groups could mechanically trap an open knot, and polymerization would create entangled knotted polymers[11].

Harry Adams*, Eleanor Ashworth*, Gloria A. Breault†, Jun Guo*, Christopher A. Hunter*, Paul C. Mayers*

**Centre for Chemical Biology, Krebs Institute for Biomolecular Science, Department of Chemistry, University of Sheffield, Sheffield S3 7HF, UK*
e-mail: c.hunter@shef.ac.uk
†AstraZeneca, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK

1. Seeman, N. C. *Angew. Chem. Int. Ed.* **37,** 3220–3228 (1998).
2. Wang, H., Di Gate, R. & Seeman, N. C. *Proc. Natl Acad. Sci. USA* **93,** 9477–9482 (1996).
3. Arai, Y. *et al. Nature* **399,** 446–448 (1999).
4. Dietrich-Buchecker, C. & Sauvage, J.-P. *Angew. Chem. Int. Ed.* **28,** 189–192 (1989).
5. Dietrich-Buchecker, C., Rapenne, G. & Sauvage, J.-P. in *Molecular Catenanes, Rotaxanes and Knots: A Journey Through the World of Molecular Topology* (eds Sauvage, J.-P. & Dietrich-Buchecker, C.) 107–142 (Wiley-VCH, Weinheim, 1999).
6. Ashton, P. R. *et al. Liebigs Ann. Recueil* **1997,** 2485–2494 (1997).
7. Safarowsky, O., Nieger, M., Fröhlich, R. & Vögtle, F. *Angew. Chem. Int. Ed.* **39,** 1616–1618 (2000).
8. Sokolov, V. I. *Russ. Chem. Rev.* **42,** 452–463 (1973).
9. Carina, R. F., Dietrich-Buchecker, C. & Sauvage, J.-P. *J. Am. Chem. Soc.* **118,** 9110–9116 (1996).
10. Breault, G. A., Hunter, C. A. & Mayers, P. C. *J. Am. Chem. Soc.* **120,** 3402–3410 (1998).
11. Breault, G. A., Hunter, C. A. & Mayers, P. C. *Tetrahedron* **55,** 5265–5293 (1999).

Neurobiology

p25 protein in neurodegeneration

The normal development of the mammalian central nervous system requires a protein kinase known as Cdk5, the activity of which depends on its interaction with a regulatory subunit, p35. An abnormal truncated form of p35 (p25) produced by the action of proteases[1,2] can also activate Cdk5, but causes apoptotic cell death in cultured primary neurons[3]. Patrick *et al.* have reported that p25 accumulates 20–40-fold in brain lysates from patients with Alzheimer's disease, and infer that p25 may contribute to the pathogenesis of neurodegeneration[3]. Here we show that the amount of p25 in the frontal cortex of patients with Alzheimer's disease or Down's syndrome is actually lower than in controls. Although there is evidence for a role for p25 in neurodegeneration from cells in culture[2,3] and in rats[4] and mice[5], our contradictory finding casts doubt on the involvement of p25

© 2001 Macmillan Magazines Ltd

in neurodegenerative conditions such as Alzheimer's disease and Down's syndrome.

We tested brains (samples from MRC London Brain Bank for Neurodegenerative Diseases) from seven patients with Down's syndrome (DS; 2 females, 5 males; mean age at death 54.57 ± 7.63 years), six with Alzheimer's (AD; 2 females, 4 males; mean age 61.33 ± 9.42 years) and eight age-matched controls with no neurological or psychiatric history (2 females, 6 males; mean age, 63.50 ± 7.31 years). The principal cause of death was bronchopneumonia for the two sets of patients and heart disease for the controls. Alzheimer samples were verified using standard criteria[6,7]. Post-mortem intervals before dissection were 35.71 ± 19.89, 26.83 ± 22.32 and 37.38 ± 20.51 hours in Down's, Alzheimer's and control brains, respectively.

We compared the levels of p25 in the different brains by probing western blots of brain homogenate (4,000*g* supernatant) with various antibodies: anti-p35 polyclonal C19 (Santa Cruz; 1:2,000), anti-p25 serum (provided by Li-Huei Tsai; 1:1,000), polyclonal N19 antibody against the protease calpain 1 (Santa Cruz; 1:100), monoclonal antibodies against β-actin (Sigma; 1:5,000) or neuron-specific enolase (NSE; Chemicon; 1:5,000).

Figure 1a shows that the intensity of the band containing p25 immunoreactive protein was similar when probed with either anti-p35 polyclonal antibody or anti-p25 antiserum. The amount of immunoreactive p25 was significantly reduced in AD ($P < 0.01$) and DS ($P < 0.05$) homogenates (Fig. 1b, c; left panels). We normalized p25 levels against β-actin and NSE to reveal relative changes in total cells and neuronal cells, respectively. The reduction was comparable in DS frontal cortex when normalized against β-actin, NSE and p35, but only the p25/NSE level was reduced in AD frontal cortex (Fig. 1b, c). In addition, levels of calpain 1 (a protease that produces p25; refs 1, 2) were comparable between groups (results not shown).

Linear-regression analysis reveals no correlation between p25 levels and age, sex, post-mortem intervals or calpain 1 protease concentration in any group (results not shown). Our finding of reduced amounts of p25 in Alzheimer's brains not only disagrees with the results of Patrick *et al.* but raises the possibility that this protein has in fact been downregulated in these patients.

Byong Chul Yoo, Gert Lubec
Department of Pediatrics, University of Vienna, 1090 Vienna, Austria
e-mail: gert.lubec@akh-wien.ac.at

1. Lee, M. *et al. Nature* **405,** 360–364 (2000).
2. Nath, R. *et al. Biochem. Biophys. Res. Commun.* **274,** 16–21 (2000).
3. Patrick, G. N. *et al. Nature* **402,** 615–622 (1999).
4. Kusakawa, G. *et al. J. Biol. Chem.* **275,** 17166–17172 (2000).
5. Ahlijanian, M. K. *et al. Proc. Natl Acad. Sci. USA* **97,** 2910–2915 (2000).
6. Tierney, M. C. *et al. Neurology* **38,** 359–364 (1988).
7. Mirra, S. S. *et al. Neurology* **41,** 479–486 (1991).

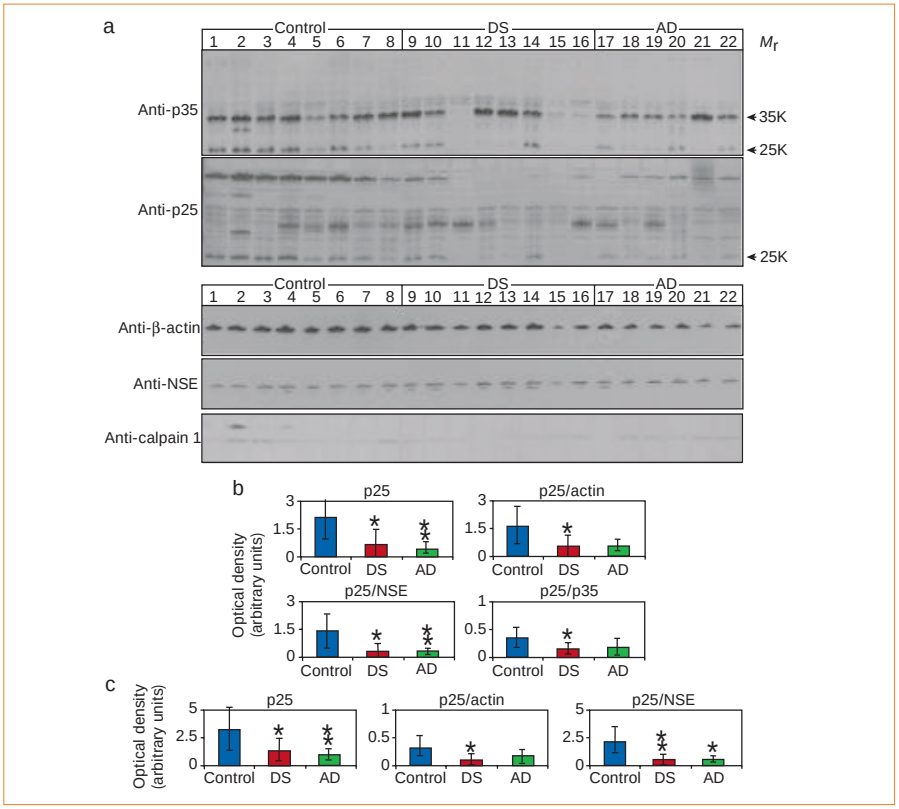


Figure 1 Reduction in the amount of p25 protein in the frontal cortex of brains from patients with Alzheimer's disease (AD) or Down's syndrome (DS). **a,** Western blots of homogenates from the frontal cortex of controls (lanes 1–8), DS (lanes 9–16; lane 15 is a fetal DS sample and data were not used for statistical analysis) and AD (lanes 17–22) probed with antibodies against p35, p25, β-actin, NSE or calpain 1. **b,** Band intensities determined by densitometry of the blots in **a** probed with anti-p35 polyclonal antibody. Levels are shown for p25 and for p25 normalized to actin, NSE or p35 intensities. All protein values are expressed as mean ± standard deviation. Between-group differences were calculated by non-parametric Mann–Whitney *U*-test; statistical significance was set at $P < 0.05$ (single and double asterisks represent $P < 0.05$ and $P < 0.01$, respectively). **c,** As **b**, but using anti-p25 serum. NSE, neuron-specific enolase.

Patrick et al. *reply* — Yoo and Lubec show that the amount of p25 is decreased in the brains they studied from patients with Alzheimer's disease or Down's syndrome. Their results persuaded us to conduct a more extensive survey of the p25/p35 ratio in AD brains (to be published elsewhere), as the number of samples was small in both of our studies (eight AD brains in our original study and six in theirs). After analysing a further 25 AD brains and those from 25 age-matched controls, we found that p25 levels are consistently higher in AD brains and that the difference is statistically significant (Student's *t*-test). This is in agreement with our original observations[1], as well as being consistent with earlier reports of increased Cdk5 kinase activity in AD brain[2] and of increased amounts of p25 in an animal model of neurodegeneration[3].

Calpain is the protease that processes p35 to generate p25. It has been shown to have increased activity in AD brains[4], but it can also be activated by hypoxic stress. p25 may therefore also be produced post mortem, which would complicate attempts to analyse human brain samples quantitatively. We found that the difference in the amounts of p25 and p35 is most marked in samples with post-mortem intervals of less than 3 hours and decreases in samples with post-mortem intervals of longer than 24 hours. These intervals average 33.3 ± 20.8 hours in the samples studied by Yoo and Lubec, which may explain the high p25 levels in their control samples.

How then can the p25 generated in AD be distinguished from p25 produced post mortem? In our latest analysis, we found p25 from AD brain to be less soluble than that from control brains. This is not surprising, given that p25 and Cdk5 are both increased in neurons containing neurofibrillary tangles. Accordingly, p25 in AD neurons is localized to less soluble compartments, unlike p25 that is generated nonspecifically during the post-mortem period.

G. N. Patrick*, L. Zukerberg*, M. Nikolic†, S. de la Monte‡, P. Dikkes§, L.-H. Tsai*
**Department of Pathology, Harvard Medical School and Howard Hughes Medical Institute, 200 Longwood Avenue, Boston, Massachusetts 02115, USA*
†Molecular Neurobiology Group, King's College London, London SE1 9RT, UK

© 2001 Macmillan Magazines Ltd

‡Rhode Island Hospital and Brown Medical School, Rhode Island 02903, USA
§Department of Neurology, Children's Hospital, Boston, Massachusetts 02115, USA

1. Patrick, G. N. *et al. Nature* **402,** 615–622 (1999).
2. Lee, K. Y. *et al. Neurosci. Res.* **34,** 21–29 (1999).
3. Nguyen, M. D., Lariviere, R. C. & Julien, J.-P. *Neuron* **30,** 135–147 (2001).
4. Grynspan, R., Griffin, W. R., Cataldo, A., Katayama, S. & Nixon, R. *Brain Res.* **763,** 145–158 (1997).

Ecosystems

Reef corals bleach to survive change

The bleaching of coral reefs, in which symbiotic algae are lost from reef-building invertebrates, is usually considered to be a drastic and damaging response to adverse environmental conditions[1,2]. Here I report results from transplant experiments involving different combinations of coral host and algal symbiont that support an alternative view, in which bleaching offers a high-risk ecological opportunity for reef corals to rid themselves rapidly of suboptimal algae and to acquire new partners. This strategy could be an advantage to coral reefs that face increasingly frequent and severe episodes of mass bleaching as a result of projected climate change[2,3].

Coral reefs are built by symbioses between scleractinian (stony) corals and photosynthetic dinoflagellate algae. These diverse algae[4] are important species because their loss during bleaching can lead to widespread coral mortality and degradation of reef ecosystems[5]. Different types of algal symbiont often show strong zonal patterns within their coral hosts that correspond to light intensity (shallow, 'high-light' algae or deep, 'low-light' algae)[6–8].

To investigate the effect of bleaching on the stability of these depth distributions, I reciprocally transplanted eight species of Caribbean scleractinian coral between 'shallow' (2–4 m) and 'deep' (20–23 m) sites in the San Blas archipelago, Panamá. I assessed transplanted and control colonies for bleaching after 8 weeks, and for mortality and changes in symbiont taxa after 12 months (Fig. 1).

'Upward' (deep-to-shallow) transplants showed significant bleaching after 8 weeks (11 of 24 colonies partially or severely bleached; 2 others pale), whereas 'downward' (shallow-to-deep) transplants showed less bleaching (0 of 37 colonies bleached; $\chi^2 = 20.7$, Fisher's exact $P < 0.0001$). Surprisingly, despite more extensive bleaching, upward transplants showed no mortality after 12 months (0 of 24 colonies dead), unlike downward transplants (7 of 37 colonies dead; $\chi^2 = 5.13$, Fisher's exact $P = 0.0358$). Control transplants showed no significant bleaching or mortality.

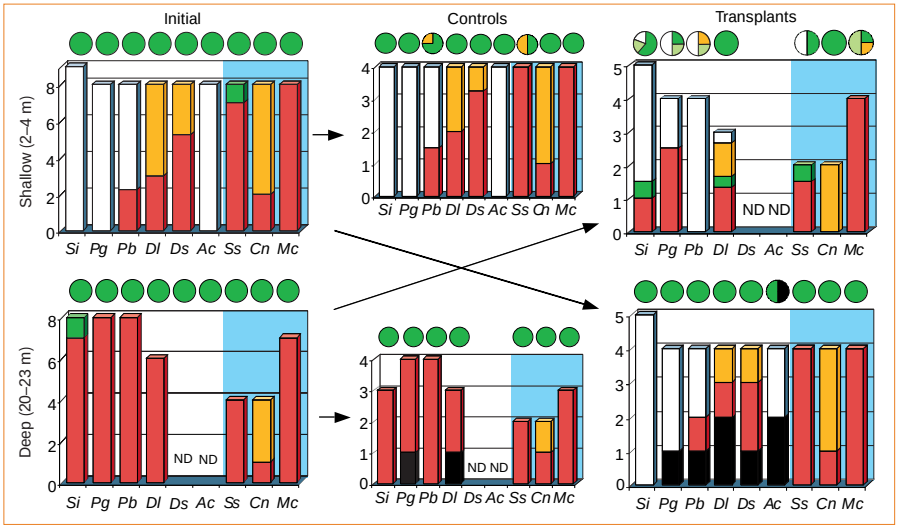


Figure 1 Symbiont diversity and mortality responses to bleaching in transplanted corals. Bars show symbiont community structure (*Symbiodinium* clades A–D) before transplantation and 12 months after transplantation: white, A; orange, B; red, C; green, D; black, dead. Pie charts indicate bleaching status of host colonies before transplantation and 8 weeks after transplantation: dark green, healthy; light green, pale; orange, partial bleaching; white, severe bleaching; black, dead. Vertical axes, number of colonies; horizontal axes, coral species. *Si, Pg, Pb, Dl, Ds* and *Ac* (white background) have strong light-related patterns of symbiont diversity; *Ss, Cn* and *Mc* (blue background) exhibit no light-related patterns of symbiont diversity. Control colonies accompanied the transplanted corals to their new location before being returned immediately to their original depths. ND, no data: *Ds* and *Ac* were very rare at the deeper location and underwent shallow-to-deep transplantation only (at similar depths nearby they were not uncommon and contained 'low-light' algae[7]). Diagonal arrows indicate direction of transplantation. *Si*, *Stephanocoenia intersepta*; *Pg*, *Porites astreoides* (green); *Pb*, *Porites astreoides* (brown); *Dl*, *Diploria labyrinthiformis*; *Ds*, *Diploria strigosa*; *Ac*, *Acropora cervicornis*; *Ss*, *Siderastrea siderea*; *Cn*, *Colpophyllia natans*; *Mc*, *Montastraea cavernosa*.

Changes in the structure of symbiont communities explained these surprising patterns of bleaching and mortality. Surveys of restriction-fragment-length polymorphisms in genes encoding large-subunit ribosomal RNA[4,8] identified four groups of *Symbiodinium* algae (termed A, B, C, and a previously unassigned clade, D[4,8,9]) from these coral hosts. Five of the eight coral host species showed strong intraspecific patterns of depth zonation in their symbionts; the other three showed no such patterns[8] (Fig. 1). Transplanted coral species that hosted different algae at deep and shallow sites adjusted their algae distributions to their new depths only when transplanted upwards (12 of 16 colonies), and not when transplanted downwards (1 of 25 colonies; $\chi^2 = 22.7$, Fisher's exact $P < 0.0001$).

These results reveal an unexpected relationship between acute stress-induced bleaching (sudden exposure to increased irradiance after upward transplantation), adaptive change in symbiont communities, and reduced coral host mortality. This contrasts with a lack of bleaching in response to chronic stress (lower sustained irradiance after downward transplantation), no change in symbiont communities, and increased coral mortality. Together, these findings support the view (first proposed by theorists[10]) that coral bleaching can promote rapid response to environmental change by facilitating compensatory change in algal symbiont communities. Without bleaching, suboptimal host–symbiont combinations persist, leading eventually to significant host mortality.

Reef corals are flexible associations that can switch or shuffle symbiont communities in response to environmental change[4,8,10,11]. However, there may be costs involved, as shown by higher mortality in the five coral species that vary their algae with depth (9 of 79 colonies) than in the three species that do not (0 of 39 colonies; $\chi^2 = 4.81$, Fisher's exact $P < 0.0289$).

Changes in symbiont communities may be slow unless existing symbionts are first removed, suggesting that established symbionts have a significant competitive home advantage over invasive (or low-abundance) symbionts. Coral bleaching can rapidly remove these symbionts, facilitating their replacement by alternative algae that are better suited to the new environmental conditions. Furthermore, the process of community change, which is facilitated by bleaching, may provide a window for unusual opportunistic symbionts to colonize hosts (and/or proliferate inside them)[9,11,12], as shown by the behaviour of *Symbiodinium* A and D in upward-transplant experiments.

Symbiosis recombination may help to resolve the paradox of reef corals as environmentally fragile yet geologically long-lived associations[13]. Despite the extreme risks involved[1,2], and the likely high incidence of mortality in some regions (such as that resulting from the 1997–98 El Niño[14]),

© 2001 Macmillan Magazines Ltd

bleaching may ultimately help reef corals to survive the recurrent and increasingly severe warming events projected by current climate models of the next half-century[3]. Bleaching is an ecological gamble in that it sacrifices short-term benefits for long-term advantage. This counters conventional wisdom that bleaching is detrimental from all perspectives, and supports the role of symbionts as adaptive agents[10,11].

Andrew C. Baker
Wildlife Conservation Society, Osborn Laboratories of Marine Science, New York Aquarium, Surf Avenue at West 8th Street, Brooklyn, New York 11224, USA
e-mail: abaker@wcs.org

1. Brown, B. E. *Coral Reefs* **16,** S129–S138 (1997).
2. Glynn, P. W. *Coral Reefs* **12,** 1–17 (1993).
3. Hoegh-Guldberg, O. *Mar. Freshwat. Res.* **50,** 839–866 (1999).
4. Rowan, R. & Powers, D. A. *Science* **251,** 1348–1351 (1991).
5. Glynn, P. W. & Colgan, M. W. *Am. Zool.* **32,** 707–718 (1992).
6. Rowan, R. & Knowlton, N. *Proc. Natl Acad. Sci. USA* **92,** 2850–2853 (1995).
7. Rowan, R., Knowlton, N., Baker, A. & Jara, J. *Nature* **388,** 265–269 (1997).
8. Baker, A. C. *The Symbiosis Ecology of Reef-Building Corals.* Thesis, Univ. Miami (1999).
9. Toller, W. W., Rowan, R. G. & Knowlton, N. *Biol. Bull. Mar. Biol. Lab. Woods Hole* (in the press).
10. Buddemeier, R. W. & Fautin, D. G. *Bioscience* **43,** 320–326 (1993).
11. Rowan, R. *J. Phycol.* **34,** 407–417 (1998).
12. Kinzie, R. A., Takayama, M., Santos, S. R. & Coffroth, M. A. *Biol. Bull. Mar. Biol. Lab. Woods Hole* **200,** 51–58 (2001).
13. Buddemeier, R. W. & Smith, S. V. *Am. Zool.* **39,** 1–9 (1999).
14. Wilkinson, C. *et al. Ambio* **28,** 188–196 (1999).

Population control

African elephants and contraception

Protected from hunting and provided with access to water-holes during droughts, elephant numbers can double in a decade, severely damaging natural vegetation and the many species dependent upon it. Culling is an effective but controversial control strategy, so Fayrer-Hosken *et al.*[1] have assessed the efficacy of using immunocontraception through vaccination, concluding that this could be a practical way of controlling elephant numbers. However, an intervention feasible in reproductive physiology may not be a practical way to control a population. Fayrer-Hosken *et al.* have not considered calculations[2,3] that undermine the practicality of their method, nor alternative management strategies.

Controlling elephants in Kruger National Park, South Africa, by immunocontraception would necessitate treatment of 2,250 cows each year over an initial period of 11 years (ref. 3). Even if individual treatments were 100% effective, the costs would be likely to exceed the total management budget of the South African national parks. The best results of Fayrer-Hosken *et al.* involved two of ten elephants becoming pregnant, and that was after receiving two booster vaccinations.

The effectiveness of this method may be less than claimed. Of the control group, 89% became pregnant within a year. This seems high, exaggerating the difference between treated and control groups. Data from 813 adult cows culled in Kruger National Park between 1979 and 1994 showed that 51% (range, 36–77%) were pregnant. This is to be expected: gestation lasts 22 months and the calving interval is 44 months (ref. 2), so about 50% of a sample of cows should be pregnant. Thus, on average, females go for 22 months without becoming pregnant. In a random sample of females monitored for 12 months, only 55% (not 89%) should therefore become pregnant.

Between 16 and 1,846 elephants of all age classes and both sexes were culled annually in Kruger National Park from 1967 to 1994. We share the desire to reduce culling and have sought methods to do so. Removing or sterilizing 250 subadult females each year should reduce population growth to zero[2,3]. Moreover, densities of greater than 0.37 elephants per square kilometre result in reduced population growth rates — probably due to reduced reproductive output by newly sexually matured females or to increased calving intervals[2]. Culling, as conducted, maintained densities at which population growth was near its maximum. Culls should be delayed for one year after counts exceed 0.37 elephants per square kilometre to allow density dependence to reduce numbers naturally[2]. Culls may still be necessary, but they would then be much less frequent and involve far fewer animals.

Stuart L. Pimm*, Rudi J. van Aarde†
**Center for Environmental Research and Conservation, Columbia University, 1200 Amsterdam Avenue, New York, New York 10027, USA*
e-mail: stuartpimm@aol.com
†Conservation Ecology Research Unit, University of Pretoria, Pretoria 0002, South Africa

1. Fayrer-Hosken, R. A., Grobler, D., Van Altena, J. J., Bertschinger, H. J. & Kirkpatrick, J. F. *Nature* **407,** 149 (2000).
2. van Aarde, R. J., Whyte, I. & Pimm, S. *Anim. Conserv.* **2,** 287–294 (1999).
3. Whyte, I., van Aarde, R. J. & Pimm, S. *Anim. Conserv.* **1,** 77–83 (1998).

Fayrer-Hosken et al. *reply* — Pimm and van Aarde question the feasibility of controlling elephant numbers by immunocontraception, arguing that the sterilization or removal of 250 subadult cows each year is the answer to population growth. However, there are no known safe methods of sterilizing free-roaming African elephants. Moving 250 subadult females to another park is impractical as there are very few areas able to receive elephants from Kruger National Park without becoming confronted with an elephant overpopulation problem of their own.

This number of subadult cows cannot be moved without disrupting the social order within their herds. Keeping them in their herds would mean that (assuming a mean herd size of 12.4, as shown in our study, and an average of 3 subadult females per herd) Kruger National Park would have to move 1,033 elephants — an unrealistic and expensive proposition. Hence the only practical way to remove 250 subadult females would be to cull them, which Pimm and van Aarde agree is an unacceptable solution.

We have shown that immunocontraception using porcine zona pellucida (pZP) works in the African elephant, although its long-term effectiveness in controlling populations is still being evaluated in South Africa. The cost and speed of field delivery have not been assessed for vaccinating large groups of elephants. However, contrary to the calculations of population modellers[1,2], immunocontraception has worked in herds of wild horses and white-tailed deer[3].

Preserving these magnificent creatures and their genetic contribution for the future is a common goal. On the basis of a single administration of a multiple-release pZP vaccine that is being developed for use in horses (I. K. M. Liu, personal communication), it should be possible to reduce the first three vaccinations used in our original study to a single dose and so minimize the stress, cost and labour of elephant immunocontraception.

We therefore question Pimm and van Aarde's criticism regarding the practicality of field immunocontraception for Kruger Park's elephant herds. It is our judgement that preserving these animals through immunocontraception is a realistic strategy that would save elephants without having to kill them.

R. A. Fayrer-Hosken*, D. Grobler, J. J. van Altena, H. J. Bertschinger, J. F. Kirkpatrick
**College of Veterinary Medicine, University of Georgia, Athens, Georgia 30602-7385, USA*
e-mail: rfh@calc.vet.uga.edu

1. van Aarde, R. J., Whyte, I. & Pimm, S. *Anim. Conserv.* **2,** 287–294 (1999).
2. Whyte, I., van Aarde, R. J. & Pimm, S. *Anim. Conserv.* **1,** 77–83 (1998).
3. Kirkpatrick, J. F., Turner, J. W., Liu, I. K. M., Fayrer-Hosken, R. A. & Rutberg, A. T. *Reprod. Fertil. Dev.* **9,** 105–110 (1997).

brief communications is intended to provide a forum for both brief, topical reports of general scientific interest and technical discussion of recently published material of particular interest to non-specialist readers. Priority will be given to contributions that have fewer than 500 words, 10 references and only one figure. Detailed guidelines are available on *Nature*'s website (www.nature.com) or on request from **nature@nature.com**

© 2001 Macmillan Magazines Ltd

The four final rotation states of Venus

Alexandre C. M. Correia & Jacques Laskar

Astronomie et Systèmes Dynamiques, IMC-CNRS UMR8028, 77 Av. Denfert-Rochereau, 75014 Paris, France

Venus rotates very slowly on its axis in a retrograde direction, opposite to that of most other bodies in the Solar System[1]. To explain this peculiar observation, it has been generally believed[2–6] that in the past its rotational axis was itself rotated to 180° as a result of core–mantle friction inside the planet, together with atmospheric tides. But such a change has to assume a high initial obliquity (the angle between the planet's equator and the plane of the orbital motion). Chaotic evolution[7], however, allows the spin axis to flip for a large set of initial conditions[6,8]. Here we show that independent of uncertainties in the models, terrestrial planets with dense atmosphere like Venus can evolve into one of only four possible rotation states. Moreover, we find that most initial conditions will drive the planet towards the configuration at present seen at Venus, albeit through two very different evolutionary paths. The first is the generally accepted view whereby the spin axis flips direction[2–6]. But we have also found that it is possible for Venus to begin with prograde rotation (the same direction as the other planets) yet then develop retrograde rotation while the obliquity goes towards zero[9]: a rotation of the spin axis is not necessary in this case.

The conservative equations for the precessional motion are[7,10]:

$$\dot{\epsilon} = A(t)\cos\psi - B(t)\sin\psi$$
$$\dot{\psi} = \alpha\cos\epsilon - 2C(t) - [A(t)\sin\psi + B(t)\cos\psi]\cot\epsilon \tag{1}$$

where $\alpha = 3/2n^2(1-e^2)^{-3/2}E_d/\omega$ is the 'precession constant' and $E_d = [(k_f R_V^5)/(3GC_V)]\omega^2 + \delta E_d$ is the dynamical ellipticity; ψ is the general precession in longitude, ϵ the obliquity, ω the inertial rotation rate, n the mean motion, e the eccentricity and G the gravitational constant. R_V is Venus' radius and $C_V > B_V > A_V$ its principal moments of inertia, k_f the fluid Love number and δE_d a residual term[11]. The quantities A, B and C depend on the secular motion of Venus' orbit[7]. The contributions of the tidal effects to Venus' spin are, at first order in e, long-term effects of the form[3,12–14]:

$$\dot{\omega} = -K^\tau \sum_\sigma b^\tau(\sigma)\Lambda_\sigma(\cos\epsilon)$$
$$\dot{\epsilon} = -K^\tau \frac{\sin\epsilon}{\omega}\sum_\sigma b^\tau(\sigma)\Theta_\sigma(\cos\epsilon) \tag{2}$$

where the superscript τ refers to the gravitational tide ($\tau = g$) or to the atmospheric tide ($\tau = a$). The functions Λ_σ and Θ_σ are polynomials in $\cos\epsilon$ (ref. 13) and σ is a tidal frequency (a linear combination of ω and n). K^τ are positive constants[3] and $b^\tau(\sigma)$ are coefficients related to the energy dissipation in the planet's interior. They depend on the time delay $\Delta t^\tau(\sigma)$ between the position of 'maximal tide' and the subsolar point as[3,13]:

$$b^g(\sigma) = k_2\sin(\sigma\Delta t^g(\sigma))$$
$$b^a(\sigma) = -\delta\tilde{p}(\sigma)\sin(\sigma\Delta t^a(\sigma)) \tag{3}$$

Here k_2 is the second potential Love number[6] and $\delta\tilde{p}(\sigma)$ the second-degree spherical harmonic component of the surface pressure variation[12]. The core–mantle friction (CMF) contribution[10,11,15–18] is given by:

$$\dot{\omega} = -\omega K^f(\omega,\kappa)\sin^2\epsilon\cos^2\epsilon$$
$$\dot{\epsilon} = -K^f(\omega,\kappa)\sin\epsilon\cos^3\epsilon \tag{4}$$

where K^f is a positive function of ω and κ, where κ is a coupling parameter that depends on the friction model.

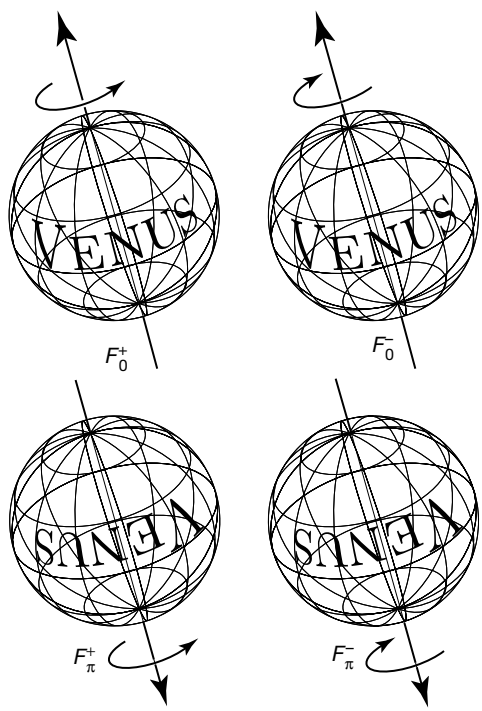
For a fast rotation rate ($\omega \gg n$), the surface pressure term $\delta\tilde{p}(\sigma)$ remains small and the contribution of atmospheric tides to the rotation rate can be neglected. The gravitational tides and the CMF both decelerate the planet's rotation, until the atmospheric tides counterbalance the braking effect and give rise to stable positions for both rotation rate and obliquity. Although many unknowns remain in the dissipative models and parameters, it is possible to show[11,16] that the CMF effect is sufficient to lead the obliquity to $\epsilon = 0°$ or $\epsilon = 180°$. This simplifies the problem, as for $\epsilon = 0°$ or $\epsilon = 180°$, the CMF contribution to $\dot{\omega}$ vanishes, and the tidal component becomes simple, with a single term of tidal frequency $\sigma = 2\omega - 2n$ for $\epsilon = 0°$ and $\sigma = 2\omega + 2n$ for $\epsilon = 180°$:

$$\dot{\omega}|_{0°} = -[K^g b^g(2\omega - 2n) + K^a b^a(2\omega - 2n)]\Lambda_{2\omega-2n}(0°)$$
$$\dot{\omega}|_{180°} = -[K^g b^g(2\omega + 2n) + K^a b^a(2\omega + 2n)]\Lambda_{2\omega+2n}(180°) \tag{5}$$

with $\Lambda_{2\omega-2n}(0°) = \Lambda_{2\omega+2n}(180°) = 3/2$. Thus, at equilibrium, with $\dot{\omega} = 0$, we obtain

$$f(\omega + \mu n) = -\frac{K^g}{K^a} \tag{6}$$

where $f(\sigma) = b^a(2\sigma)/b^g(2\sigma)$, $\mu = -1$ for $\epsilon = 0°$ and $\mu = +1$ for $\epsilon = 180°$. The invariance in equations (2) when changing ω to $-\omega$ and ϵ to $\epsilon + \pi$ leads $f(\sigma)$ to be an even function of σ. For all



State	ϵ	ω	P (d)
F_0^+	0°	$n + \omega_s$	76.83
F_0^-	0°	$n - \omega_s$	−243.02
F_π^+	180°	$-n - \omega_s$	−76.83
F_π^-	180°	$-n + \omega_s$	243.02

Figure 1 Possible final spin states of Venus. There are two retrograde states (F_0^- and F_π^-) and two direct states (F_0^+ and F_π^+). With $\omega_{obs} = 2\pi/243.0185$ d and $n = 2\pi/224.701$ d, the synodic period corresponds to $\omega_s = 2\pi/116.751$ d.

© 2001 Macmillan Magazines Ltd

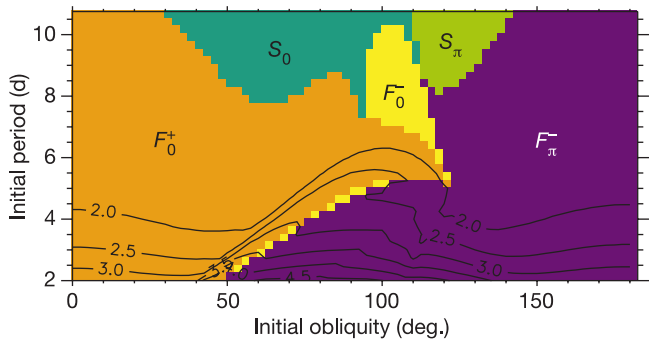


Figure 2 Numerical simulations of the final states of Venus' spin in the absence of planetary perturbations. Calculations were performed for initial obliquity ϵ_0 in the range 0° to 180°, and initial period P_0 in the range 2 d to 10.5 d. F_π^+ is not attainable, and only three of the four possible final states (Fig. 1) are reached. F_0^+ is the only direct state (orange): the spin axis is tilted to 0° and the rotation rate slows down until it reaches the final speed $\omega = n + \omega_s$ (76.83 d). When the initial period is larger than 5 d, the planet can be captured in the 1 : 1 resonance ($S_0 : \omega = n, \epsilon = 0°$) before the atmosphere begins to exert an effect (dark green). For high initial obliquities, the spin axis is reversed and the rotation rate slows down until it stabilizes in the retrograde state F_π^- (violet). Here also, for high initial periods, the planet can be captured in the 1 : 1 resonance ($S_\pi : \omega = -n, \epsilon = 180°$) (light green). The third possibility is the retrograde state F_0^- (yellow). As F_π^-, this final state corresponds to Venus' present observed state, but with a completely different history: the obliquity is decreased to 0° and the rotation speed reduces to zero. it is then accelerated in the other direction and stabilizes at $\omega = -n + \omega_s$ (243.02 d). The time after which the final state is reached is indicated by curves labelled in Gyr.

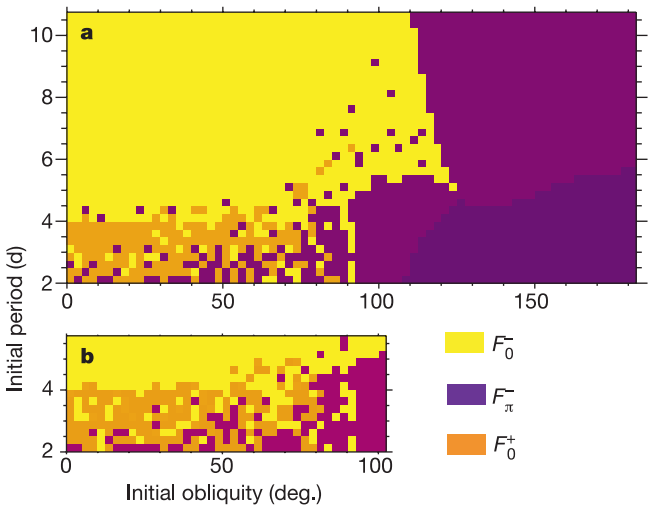


Figure 3 Numerical simulations of the final states of Venus' spin when planetary perturbations are included. (The ranges of initial obliquity and period are as in Fig. 2.) **a**, The resonant zones become unstable, and the probability of capture in a resonance becomes zero. We also observe a reduction of the F_0^+ zone. The passage through the chaotic zone is reflected by the scattering of the final states in the bottom left corner. To emphasize this chaotic behaviour, we integrated this area a second time with the same initial conditions, but with a difference of 10^{-9} in the initial eccentricity of Mars (**b**). As expected, although the overall aspect of this small area is the same, the distribution of the final states is different to that in **a**.

dissipative models[2,5,19,20], the restriction $\tilde{f}$ of f to $\sigma \geq 0$ is decreasing for slow rotation rates. There are thus only four possible values for the final rotation rate ω_f of Venus, given by:

$$|\omega_f + \mu n| = \tilde{f}^{-1}\left(-\frac{K^g}{K^a}\right) \equiv \omega_s \qquad (7)$$

Assuming that the observed rotation state of Venus corresponds to a stable retrograde rotation[11,19,21], the only possibilities for this observed final state are $\epsilon = 0°$ and $\omega_{obs} = n - \omega_s$, or $\epsilon = 180°$ and $\omega_{obs} = \omega_s - n$. In both cases, $\omega_s = n + |\omega_{obs}|$, and we can determine all four final states for Venus (Fig. 1). The two retrograde states (F_0^- and F_π^-) correspond to the present spin of Venus with period 243.02 d, while the direct states (F_0^+ and F_π^+) have a rotation period of 76.83 d. We note that these final states do not depend on the precise dissipative model and will persist for a wide range of parameters.

In our numerical experiments, we had to be more specific. For the gravitational tides, we used a simple interpolation between the constant model (for high tidal frequencies) and the linear model (for small tidal frequencies)[3]:

$$b^g(\sigma) = k_2 \frac{\mathrm{sign}(\sigma)}{Q}\left(1 - (1 - Q/Q_n)^{\frac{|\sigma|}{n}}\right)$$

where $\mathrm{sign}(\sigma) = +1$ if $\sigma \geq 0$ and -1 if $\sigma < 0$, Q is the quality factor for a fast-rotating planet and Q_n the same factor for the tidal frequency $\sigma = n$. We used $Q = 21.5$ and $Q_n = 50$ (refs 11, 22). $b^a(\sigma)$ is computed assuming that $\Delta t^a(\sigma)/\Delta t^g(\sigma) = \Delta t^a(2\omega_s)/\Delta t^g(2\omega_s) \approx k_2 K^g/\delta\tilde{p}(2\omega_s)K^a$ (equations (6) and (7)), and using the 'heating at the ground' model[20]. For CMF we used Rochester's model for laminar friction[15,17]. We chose a strong viscosity $\nu = 10^{-2}\,\mathrm{m^2\,s^{-1}}$ that delays the onset of the turbulence. Once in the turbulent regime, we assumed that $\kappa \gg e_c\omega$ (where e_c is the core ellipticity), which is equivalent to neglecting the inertial pressure couple[8,18].

Planetary atmospheres result from an evolutionary process that takes several hundred million years[23–26]. Therefore, in our numerical simulations the effect of a dense atmosphere is only included after 1 Gyr. In a first experiment, we numerically integrated the system without considering the effects of planetary perturbations (Fig. 2), and we selected initial periods that led to a final state in less than 4.6 Gyr.

Figure 3 is the same as Fig. 2 but with the introduction of the planetary perturbations[27]. If during its dynamical history, Venus' obliquity crosses the large chaotic zone that lies between 0° and 90°, its final evolution will be largely modified, whereas for initial obliquity greater than 90°, it remains essentially unchanged[7].

The most notable and unexpected difference between Fig. 2 and Fig. 3 is the replacement of nearly all the F_0^+ zone (orange) by a F_0^- zone (yellow). In Fig. 2, the direct final state F_0^+ was reached largely because, before the formation of a dense atmosphere, the spin axis was quickly set to $\epsilon = 0°$ and the rotation rate slowed down towards the synchronous period with $\omega > n$. As the atmosphere's contribution increases, the spin rate is accelerated towards $\omega = n + \omega_s$ (final state F_0^+). In the presence of planetary perturbations, the obliquity never goes to zero[11] ($\bar{\epsilon} \approx 2°$) and the CMF effect drives the spin rate towards $\omega/n \approx 1 - nQ_nK^f(n,\kappa)\sin^2\epsilon/(3K^g k_2)$. This same effect prevents the planet from being captured in the 1 : 1 resonance. Indeed, when $\omega \approx n$, the CMF torque can be decomposed into a negative constant component, and a positive term proportional to $(\omega/n - 1)$, which are each of the same magnitude as the gravitational tidal torque. Because $(B_V - A_V)/C_V \approx 2.2 \times 10^{-6}$ (ref. 28), we found that the probability of capture in resonance is zero[29]. Thus, once the atmosphere contribution becomes important, the rotation rate is driven towards $\omega = n - \omega_s$ (final state F_0^-). For weaker CMF effects, the capture into resonance is possible, though with small probability. However, when the atmosphere effect becomes important, the planet will inevitably leave the resonance. Finally, for initial period smaller than 5 d, and obliquity smaller than 90°, we have a small zone where all three final states (F_0^+, F_0^-, F_π^-) are distributed randomly (Fig. 3), which results from the crossing of the chaotic zone[7].

To estimate the probability of each final state in this zone, we integrated 100 cases with very close initial conditions, for every

© 2001 Macmillan Magazines Ltd

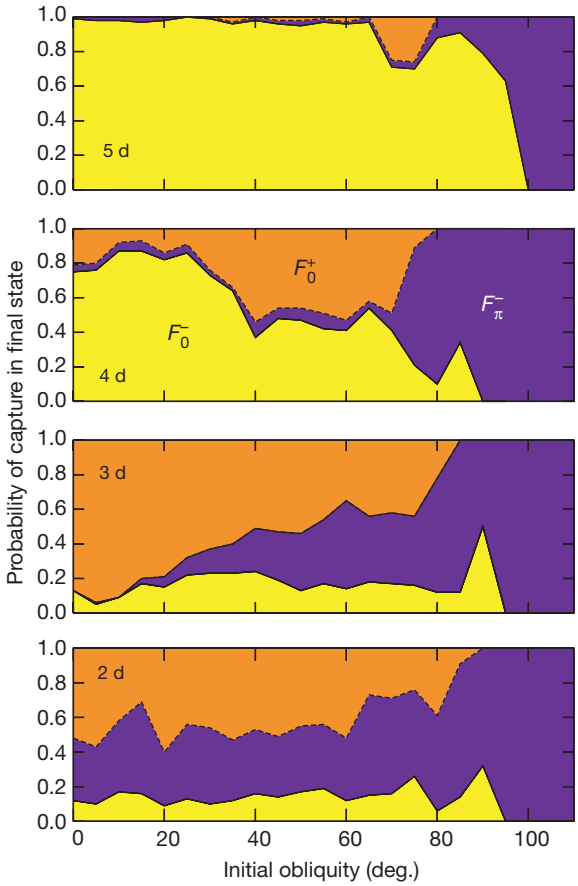


Figure 4 Probability of capture in one of Venus' final states: orange, F_0^+; yellow, F_0^-; violet, F_π^-. For initial periods of 2, 3, 4 and 5 d (bottom to top, respectively) and initial obliquity from 0° to 110°, with 5° step size, 100 trajectories were computed with an initial increment of 10^{-6} radian in their initial precession angle.

initial obliquity and initial periods P_0 of 2, 3, 4 and 5 d (Fig. 4). Except for an initial rotation period of 3 d and low obliquity, most of the initial conditions lead to the present state of Venus (F_0^- or F_π^-). For a rapid initial rotation speed ($P_0 \approx 2$ d), F_π^- is the most probable, but for $P_0 \geq 4$ d, and for any initial obliquity smaller than 70°, the most probable final state is F_0^-.

These probabilities depend on many unknowns, but the general behaviour depicted here should be robust. A different value for the dissipation factor Q will change the vertical axis values of the initial periods in Fig. 2, but should not change the displayed results much. Similarly, we have tested a CMF model with smaller viscosity (like the one used by Yoder[11]), with no significant changes (data not shown).

Our results are not modified by a linear increase of the atmosphere density from 500 Myr to 1 Gyr. If the atmosphere is introduced earlier, between 300 Myr and 800 Myr (refs 24–26), the results of Fig. 3 will be slightly displaced—the zone where direct states dominate will be extended up to $P_0 = 5.5$ d instead of $P_0 = 4$ d. In the extreme case where the dense atmosphere is present from the very beginning, the near synchronization states would never be attained. The yellow zone of Fig. 3 would then be transformed to a zone of mixed states (F_0^+, F_0^-, F_π^-), similar to the behaviour obtained at present for $P_0 = 3$ d (Fig. 4), where direct final states are prevailing, leaving less chances of ending up in the present configuration.

The present study confirms that the retrograde rotation of Venus can be explained by a consideration of its dynamics, starting with any initial obliquity or period. The key result is that this retrograde state is not necessarily obtained through the F_π^- state, but can also be obtained through the F_0^- state. The existence of the four final states (F_0^+, F_π^+, F_0^-, F_π^-) that are depicted here is a very general feature of the rotational evolution of a planet with a dense atmosphere, and should be of interest to the dynamical analysis of extrasolar planets. In particular, we expect that extrasolar planets in the F_0^+ state will be observed. □

Received 8 August 2000; accepted 17 April 2001.

1. Carpenter, R. L. A radar determination of the rotation of Venus. *Astron. J.* **75,** 61–66 (1970).
2. Lago, B. & Cazenave, A. Possible dynamical evolution of the rotation of Venus since formation. *Moon Planets* **21,** 127–154 (1979).
3. Dobrovolskis, A. R. Atmospheric tides and the rotation of Venus. II—Spin evolution. *Icarus* **41,** 18–35 (1980).
4. Shen, M. & Zhang, C. Z. Dynamical evolution of the rotation of Venus. *Earth Moon Planets* **43,** 275–287 (1989).
5. McCue, J. & Dormand, J. R. Evolution of the spin of Venus. *Earth Moon Planets* **63,** 209–225 (1993).
6. Yoder, C. F. *Venus II: Geology, Geophysics, Atmosphere, and Solar Wind Environment* 1087 (Univ. Arizona Press, Tucson, 1997).
7. Laskar, J. & Robutel, P. The chaotic obliquity of the planets. *Nature* **361,** 608–612 (1993).
8. Néron de Surgy, O. *Influence des Effets Dissipatifs sur les Variations à Long Terme des Obliquités Planétaires*. Thesis, Observatoire de Paris (1996).
9. Kundt, W. Spin and atmospheric tides of Venus. *Astron. Astrophys.* **60,** 85–91 (1977).
10. Néron de Surgy, O. & Laskar, J. On the long term evolution of the spin of the Earth. *Astron. Astrophys.* **318,** 975–989 (1997).
11. Yoder, C. F. Venus' free obliquity. *Icarus* **117,** 250–286 (1995).
12. Chapman, S. & Lindzen, R. *Atmospheric Tides. Thermal and Gravitational* (Reidel, Dordrecht, 1970).
13. Kaula, W. Tidal dissipation by solid friction and the resulting orbital evolution. *Rev. Geophys.* **2,** 661–685 (1964).
14. Munk, W. H. & MacDonald, G. J. F. *The Rotation of the Earth, a Geophysical Discussion* (Cambridge Univ. Press, Cambridge, 1960).
15. Rochester, M. G. The secular decrease of obliquity due to dissipative core-mantle coupling. *Geophys. J. R. Astron. Soc.* **46,** 109–126 (1976).
16. Goldreich, P. & Peale, S. J. The obliquity of Venus. *Astron. J.* **75,** 273–284 (1970).
17. Pais, M. A. *et al.* Late Precambrian paradoxical glaciation and obliquity of the Earth—a discussion of dynamical constraints. *Earth Planet. Sci. Lett.* **174,** 155–171 (1999).
18. Aoki, S. Friction between mantle and core of the Earth as a cause of the secular change in obliquity. *Astron. J.* **74,** 284–291 (1969).

© 2001 Macmillan Magazines Ltd

19. Gold, T. & Soter, S. Atmospheric tides and the resonant rotation of Venus. *Icarus* **11,** 356–366 (1969).
20. Dobrovolskis, A. R. & Ingersoll, A. P. Atmospheric tides and the rotation of Venus. I—tidal theory and the balance of torques. *Icarus* **41,** 1–17 (1980).
21. Davies, M. E. *et al.* The rotation period, direction of the north pole, and geodetic control network of Venus. *J. Geophys. Res.* **97,** 13141–13151 (1992).
22. Dickey, J. O. Lunar laser ranging—a continuing legacy of the Apollo program. *Science* **265,** 482–490 (1994).
23. Walker, J. C. G. Evolution of the atmosphere of Venus. *J. Atmos. Sci.* **32,** 1248–1256 (1975).
24. Pepin, R. O. On the origin and early evolution of terrestrial planet atmospheres and meteoritic volatiles. *Icarus* **92,** 2–79 (1991).
25. Hunten, D. M. Atmosphere evolution of the terrestrial planets. *Science* **259,** 915–920 (1993).
26. Kasting, J. F. Earth's early atmosphere. *Science* **259,** 920–926 (1993).
27. Laskar, J. The chaotic motion of the solar system. A numerical estimate of the size of the chaotic zones. *Icarus* **88,** 266–291 (1990).
28. Konopliv, A. S. *et al.* Venus gravity and topography: 60th degree and order model. *Geophys. Res. Lett.* **20,** 2403–2406 (1993).
29. Goldreich, P. & Peale, S. Spin-orbit coupling in the solar system. *Astron. J.* **71,** 425–438 (1966).

Acknowledgements

We thank O. Néron de Surgy, M. Greff, S. Peale and C. F. Yoder for discussions. This work was supported by PNP-CNRS and by the Fundação para a Ciência e Tecnologia, Portugal.

Correspondence and requests for materials should be addressed to J.L. (e-mail: laskar@bdl.fr).

Persistent sourcing of coherent spins for multifunctional semiconductor spintronics

I. Malajovich*, J. J. Berry†, N. Samarth† & D. D. Awschalom*

* *Department of Physics, University of California, Santa Barbara, California 93106, USA*

† *Department of Physics, Pennsylvania State University, University Park, Pennsylvania 16802, USA*

Recent studies of n-type semiconductors have demonstrated spin-coherent transport over macroscopic distances[1], with spin-coherence times exceeding 100 ns[2,3]; such materials are therefore potentially useful building blocks for spin-polarized electronics ('spintronics'). Spin injection into a semiconductor (a necessary step for spin electronics[4]) has proved difficult[5,6]; the only successful approach involves classical injection of spins from magnetic semiconductors[7,8]. Other work has shown that optical excitation can provide a short (<500 ps) non-equilibrium burst of coherent spin transfer across a GaAs/ZnSe interface, but less than 10% of the total spin crosses into the ZnSe layer, leaving long-lived spins trapped in the GaAs layer (ref. 9). Here we report a 'persistent' spin-conduction mode in biased semiconductor heterostructures, in which the sourcing of coherent spin transfer lasts at least 1–2 orders of magnitude longer than in unbiased structures. We use time-resolved Kerr spectroscopy to distinguish several parallel channels of interlayer spin-coherent injection. The relative increase in spin-coherent injection is up to 500% in the biased structures, and up to 4,000% when p–n junctions are used to impose a built-in bias. These experiments reveal promising opportunities for multifunctional spin electronic devices (such as spin transistors that combine memory and logic functions), in which the amplitude and phase of the net spin current are controlled by either electrical or magnetic fields.

GaAs/ZnSe samples are designed for measurements of interlayer spin transport under electrical bias. Figure 1a shows this heterostructure together with a proposed band structure[10]. The 400-μm-thick GaAs substrates are n-doped with Si concentrations of 3×10^{16} cm^{-3} and back-contacted with indium. ZnSe epilayers, grown on the substrates by molecular beam epitaxy, are n-doped with Cl donors and transparently contacted by 40 nm of indium tin oxide[11]. The epilayers are 100, 150, 200 and 300 nm thick and have a carrier density of 5–15 $\times 10^{17}$ cm^{-3}; for brevity, only results from two extreme samples (100 nm thick, $n = 5 \times 10^{17}$ cm^{-3} and 300 nm thick, $n = 15 \times 10^{17}$ cm^{-3}) will be presented.

The two-colour pump-probe measurement strategy is described in detail elsewhere[9]. Briefly, a 1.5-eV, 100-fs circularly polarized laser pulse selectively excites polarized conduction band spins in GaAs. At a time Δt later, the spin polarization in ZnSe is measured by a 2.8-eV linearly polarized probe pulse, whose polarization rotates by an amount proportional to the spin magnetization in the propagation direction (Kerr effect). Figure 1b (top) illustrates the spin dynamics in the absence of electrical bias, where between 2 and 10% of the spins in GaAs spontaneously transfer to the epilayer during the first few hundred picoseconds[9]. Most of the spins remain trapped in the substrate, which acts as a spin reservoir owing to its long spin lifetime[2]. Here, we discuss an additional mode of spin injection that arises under an external bias, wherein a persistent spin current is sourced by the reservoir (Fig. 1b, bottom).

Figure 1c shows that the electric and magnetic responses of the spin current are not independent in biased structures. Previous studies have shown that the spin character of a homogeneous material (lifetime and g-factor) is revealed by sweeping an external

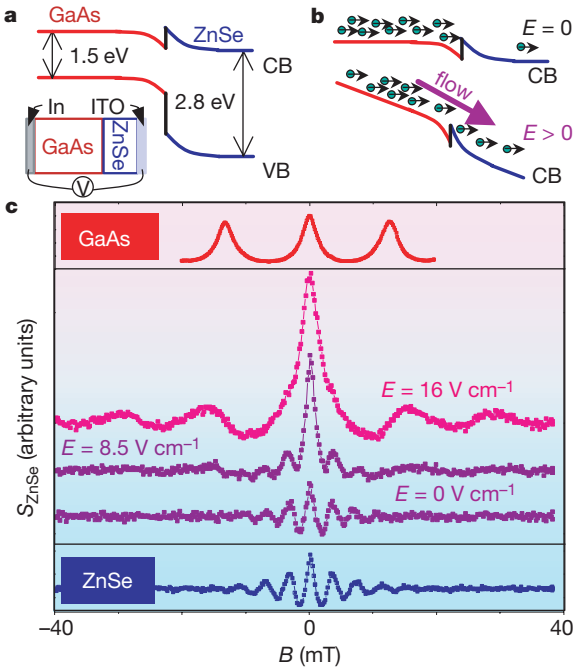


Figure 1 Magnetic field response of the transferred spins evolving from ZnSe-like to GaAs-like with an electrical bias. **a**, Schematic conduction band (CB) and valence band (VB) diagram of the n-GaAs/n-ZnSe heterostructure. The offset between the different layers is not known, because the band bending and interface potential will depend on growth details. **b**, Schematic representation of the spin transfer in the conduction band with (bottom) and without (top) bias. **c**, Resonance scans taken by sweeping the applied magnetic field at a fixed delay of −10 ps at $T = 5$ K in a 100-nm-thick $n = 5 \times 10^{17}$ cm^{-3} epilayer. The top red (bottom blue) spectrum corresponds to the pump and probe pulses tuned to the GaAs (ZnSe) band gap. For the GaAs spectrum a 400-W cm^{-2} pump and a 40-W cm^{-2} probe are used; the ZnSe spectrum is taken with 80 W cm^{-2} and 30 W cm^{-2} for the pump and probe, respectively. The measurement of spins transferred from the GaAs substrate to the ZnSe epilayer at different biases is shown (middle). The response of the transferred spins to a magnetic field change from that characteristic of ZnSe to that characteristic of GaAs with increasing bias, as indicated by the colour of each curve and the graded background. All the interface transfer data is taken using a 200-W cm^{-2}, 1.52-eV pump and a 30-W cm^{-2}, 2.79-eV probe.

© 2001 Macmillan Magazines Ltd

magnetic field B at fixed pump-probe delay, Δt (ref. 2). Periodicities in the pump-pulse train yield resonances by spin interference, whose spacing and linewidth give a measure of spin g-factors and transverse lifetimes, respectively[2,12]. Figure 1c shows the resonance spectra for GaAs (top, red) and ZnSe (bottom, blue) layers, obtained by exciting and detecting spins in the same layer (degenerate pump probe). Also shown are resonance spectra obtained by exciting in the GaAs layer and detecting spins that have crossed into the ZnSe epilayer (two-colour pump probe) under different bias conditions. We note that as the bias increases, the resonance spectrum transforms from that characteristic of ZnSe to one whose spacing more closely resembles that of GaAs. The data show that an electric field can be used to control the magnetic response of the ZnSe layer during spin injection, and vice versa. As we will show below, this unusual multifunctional behaviour results from concurrent changes in the amplitude and orientation of the spin current itself.

Time-resolved measurements show that such changes arise from the onset of a fundamentally new mode of spin transfer. Without an external bias (Fig. 2a, b for $E = 0$), spin accumulation in the epilayer lasts only a few hundred picoseconds and can leave a large spin polarization trapped in the reservoir[9]. Once transferred, spins precess about B at a rate determined by the ZnSe g-factor, and decay with a time reflecting the ZnSe spin lifetime. The following changes are observed in the spin polarization transfer upon biasing (Fig. 2a, b for $E > 0$): (1) the amplitude of the spin transfer increases, (2) the spin signal survives for longer times, and (3) a second precession frequency corresponding to a different g-factor appears. In epilayers with lifetimes longer than 4 ns, the appearance of a second g-factor is revealed by the presence of beats (Fig. 2a); in contrast, in epilayers with shorter spin lifetimes (Fig. 2b), after a few nanoseconds the spin signal is present only under a biased condition and exhibits a precession rate different from that of ZnSe. A Fourier transform of the biased time scans shows that the second g-factor is within 2% of the nominal −0.44 GaAs value.

These data indicate that when $E > 0$, spins trapped in the GaAs spin reservoir are pulled into the ZnSe epilayer, continuously sourcing a spin current. These concepts are made more quantitative by adding new spin conduction channels to the transient spin transfer mechanism that describes unbiased spin injection[9]. In that model, the epilayer spin is

$$S_{\mathrm{ZnSe}}(\Delta t) = A\sum_{n=0}^{\infty}\left\{e^{-\frac{\Delta t + nt_{\mathrm{rep}}}{T^*_{2\mathrm{ZnSe}}}}\cos(\omega_{\mathrm{ZnSe}}\Delta t + \phi) - e^{-\frac{\Delta t + nt_{\mathrm{rep}}}{\tau_{\mathrm{eff}}}}\cos(\omega_{\mathrm{GaAs}}\Delta t + \phi)\right\} \quad (1)$$

where $\phi = \tan^{-1}(\tau_{\mathrm{eff}}(\omega_{\mathrm{GaAs}} - \omega_{\mathrm{ZnSe}}))$, $T^*_{2\mathrm{GaAs}}$ and $T^*_{2\mathrm{ZnSe}}$ are the GaAs and ZnSe spin lifetimes, respectively, ω_{GaAs} and ω_{ZnSe} their Larmor frequencies (which contain the g-factor explicitly), t_{rep} is the laser repetition time, $\tau_{\mathrm{eff}}^{-1} = \tau^{-1} + T^{*-1}_{2\mathrm{GaAs}}$ is the effective spin accumulation time and τ is the spin accumulation time. This equation is obtained by summing over spins crossing the interface at different times t_{i} with an exponentially decreasing probability $\propto e^{-t_{\mathrm{i}}/\tau}$ (the exact form of ref. 9 is recovered if $\tau \ll T^*_{2\mathrm{ZnSe}}$ and if $T^*_{2\mathrm{ZnSe}} < t_{\mathrm{rep}}$). Here, we refine the modelling of the unbiased data by considering the sum of parallel fast and slow transfer mechanisms, A and B, each described by functions S^{A} and S^{B} of the form of equation (1) but with amplitudes A and B, respectively, and corresponding accumulation times $\tau^A_{\mathrm{eff}} \approx \tau^{\mathrm{A}} \approx 20$ ps and $\tau^B_{\mathrm{eff}} \approx \tau^{\mathrm{B}} \approx 500$ ps (where we used $T^*_{2\mathrm{GaAs}} \gg \tau^{A,B}$). Long spin lifetimes require summation over prior pump pulses as well. To fit the biased spin transfer data, the addition of a persistent spin flow (mechanism C) is needed, described by equation (1) with an amplitude C and $\tau^C = \infty$. In this case spin transfer never turns off, and $\tau^C_{\mathrm{eff}} \approx T^*_{2\mathrm{GaAs}}$ represents not the spin accumulation time but a decay in the persistent spin current that mirrors spin polarization decay in the reservoir. The total spin signal in the epilayer is therefore described by $S_{\mathrm{ZnSe}}(\Delta t) = S^A + S^B + S^C$, used as fits in Fig. 2a, b. Figure 2c shows the dynamics of the three parallel channels contributing to spin injection: the spins that cross over tens of picoseconds (S^A), over a few hundred picoseconds (S^B),

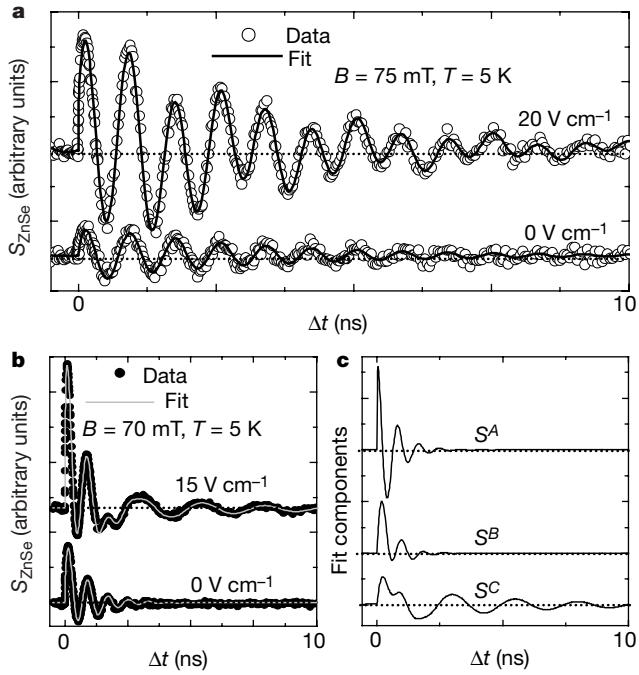


Figure 2 Time progression of the spin transfer from the reservoir to the epilayer with and without a bias. An offset is added for clarity; the dotted lines mark the zeros. **a**, 300-nm-thick, $n = 1.5 \times 10^{18}$ cm^{-3} ZnSe epilayer. **b**, 100-nm-thick, $n = 5 \times 10^{17}$ cm^{-3} ZnSe epilayer. **c**, Decomposition of the fit to the biased scan in **b** into its three different conduction channels.

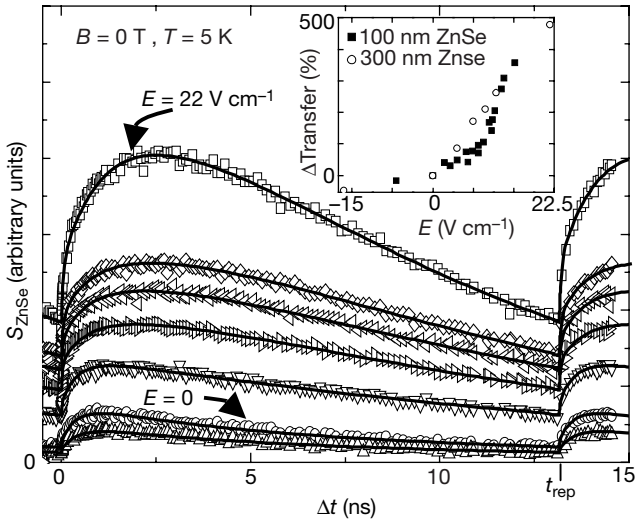


Figure 3 Changes in the efficiency of the spin transfer with bias. Time evolution of the spin transfer at different biases with $B = 0$ T. The time interval is long enough to include two successive pump pulses. The 300-nm-thick, $n = 1.5 \times 10^{18}$ cm^{-3} ZnSe epilayer was used with $E = -16.5$, 0, 4.5, 7.5, 9.8, 12 and 22 V cm^{-1} (at higher voltages heating occurred). Solid line indicates fit to the three-channel model. No offset added. Inset, percentile change in the spin transfer from the reservoir to the epilayer, calculated using the sum of the three mechanisms amplitudes ($A + B + C$) as a measure of the spin transfer at different biases, and comparing them to the unbiased case.

© 2001 Macmillan Magazines Ltd

and the persistent spin current that appears with a bias (S^C). Fitting variables are the amplitudes A, B and C (C is taken as non-zero only for positive biases), and the risetimes τ^A and τ^B. The other values are fixed and obtained from complementary measurements, such as resonance spectra or time scans from degenerate pump-probe arrangements. We note that the persistent spin current contribution is sensitive only to spin that has arrived within a time $T^*_{2\mathrm{ZnSe}}$. Hence, for $t > T^*_{2\mathrm{ZnSe}}$, the epilayer polarization reflects the precession and the lifetime of the spin current itself, which follows the reservoir spin dynamics. This explains why the average spin polarization in the epilayer is characterized by ω_{GaAs} and $T^*_{2\mathrm{GaAs}}$ although spins reside in ZnSe (and each spin has ω_{ZnSe} and $T^*_{2\mathrm{ZnSe}}$).

The changes in net spin transfer with bias are more evident in the absence of spin precession ($B = 0$). In Fig. 3, time scans show an increase in the amplitude of the spin signal in ZnSe as the applied bias is increased, while a reverse bias reduces the number of spins that cross the interface. The most striking influence of the persistent spin flow (C) on spin transfer is to introduce an apparent offset to the spin polarization. This offset is due to the long spin lifetimes in GaAs: the reservoir spin population decays only slightly before receiving a boost from the next pump pulse, sourcing a spin current that never reaches zero for $E > 0$. The inset in Fig. 3 shows the change in the total spin polarization transferred from the reservoir to the epilayer, where modest electric fields increase spin transfer nearly 500%. About 3/4 of this increase is due to the persistent current; mechanisms A and B, also enhanced by the bias, account for the rest. We also find that the change in spin transfer mimics changes in charge current across the structure (from I–V measurements), showing that spin and charge motion are coupled in mechanism C.

To see whether similar physics arises from the built-in interfacial electric fields in a p–n junction, we also study spin transfer across a p-GaAs/n-ZnSe heterojunction. As a control, an undoped GaAs substrate is placed together with the p-type substrate in the growth chamber, and 300 nm of n-ZnSe ($n = 1.5 \times 10^{18}\ \mathrm{cm}^{-3}$) is deposited simultaneously on both. Figure 4a shows a roughly 4,000% increase in spin transfer due to the heterojunction voltage, compared with the control. Degenerate pump-probe measurements of the ZnSe epilayers are nearly identical, with a slight change in spin lifetime (Fig. 4b). Therefore, the difference in the transferred spin signal amplitudes between the two samples shown in Fig. 4a is not due to differences in the ZnSe epilayers themselves. Unlike the n-doped GaAs 'spin reservoirs', the electron spin lifetimes in the p-doped and undoped GaAs substrates are extremely short[2] (a few picoseconds) so that persistent spin currents do not explain the observed increase. Instead, the data suggest enhancement of spontaneous transfer mechanisms similar to A and B, but with a non-exponential spin accumulation profile. This is consistent with the observation of electron spin precession at a single frequency corresponding to the g-factor of ZnSe rather than GaAs.

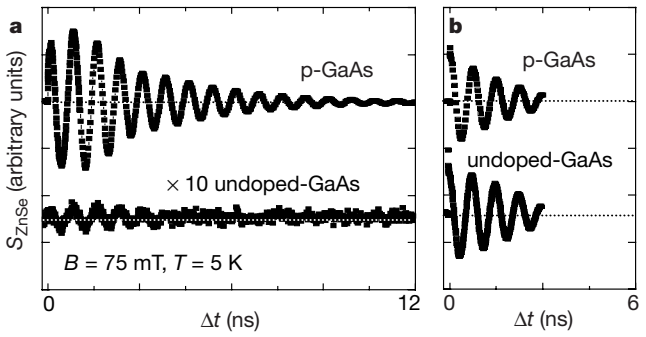


Figure 4 Time evolution of the spin polarization transfer across a p-GaAs/n-ZnSe heterojunction. Zn p-type doping concentration, $10^{19}\ \mathrm{cm}^{-3}$. Transfer across an undoped-GaAs/n-ZnSe is shown for comparison. Dotted lines show the zeros, and an offset was added for clarity. **a**, Spins transported across the interface (two-colour pump probe). **b**, Spins excited and measured in the ZnSe epilayer (degenerate pump probe).

In summary, we show that persistent spin currents appear in biased spin injection studies of GaAs/ZnSe heterostructures. We find that—under electrical bias—the GaAs layer behaves as a coherently rotating spin reservoir that continuously sources current to the ZnSe layer. As a result, the spin transfer efficiency is increased markedly. Furthermore, the magnetic and electric responses become interdependent, thus providing a new paradigm for the future development of magneto-electronic semiconductor devices with qualitatively new functionality. □

Received 28 February; accepted 2 April 2001.

1. Kikkawa, J. M. & Awschalom, D. D. Lateral drag of spin coherence in gallium arsenide. *Nature* **397,** 139–141 (1999).
2. Kikkawa, J. M., Smorchkova, I. P., Samarth, N. & Awschalom, D. D. Room-temperature spin memory in two-dimensional electron gases. *Science* **277,** 1284–1287 (1997).
3. Kikkawa, J. M. & Awschalom, D. D. Resonant spin amplification in n-type GaAs. *Phys. Rev. Lett.* **80,** 4313–4316 (1998).
4. Prinz, G. A. Spin polarized transport. *Phys. Today* **48,** 58–63 (1995).
5. Monzon, F. G. & Roukes, M. L. Spin injection and the local hall effect in InAs quantum wells. *J. Mag. Magn. Mater.* **198,** 632–635 (1999).
6. Filip, A. T., Hoving, B. H., Jedema, F. J. & van Wees, B. J. Experimental search for the electrical spin injection in a semiconductor. *Phys. Rev. B* **62,** 9996–9999 (2000).
7. Fiederling, R. *et al.* Injection and detection of a spin-polarized current in a light-emitting diode. *Nature* **402,** 787–790 (1999).
8. Ohno, Y. *et al.* Electrical spin injection in a ferromagnetic semiconductor heterostructure. *Nature* **402,** 790–792 (1999).
9. Malajovich, I., Kikkawa, J. M., Awschalom, D. D., Berry, J. J. & Samarth, N. Coherent transfer of spin through a semiconductor heterointerface. *Phys. Rev. Lett.* **84,** 1015–1018 (2000).
10. Yeganeh, M. S., Qi, J., Yodh, A. G. & Tamargo, M. C. Interface quantum well states observed by three-wave mixing in ZnSe/GaAs heterostructures. *Phys. Rev. Lett.* **68,** 3761–3764 (1992).
11. Schull, K. *et al.* Non-metal in situ and ex situ ohmic contacts to a n-ZnSe. *Semicond. Sci. Technol.* **12,** 485–489 (1997).
12. Malajovich, I., Kikkawa, J. M., Awschalom, D. D., Berry, J. J. & Samarth, N. Resonant amplification of spin transferred across a GaAs/ZnSe interface. *J. Appl. Phys.* **87,** 5073–5075 (2000).

Acknowledgements

We thank M. E. Flatte, E. L. Hu, J. M. Kikkawa and H. Kroemer for discussions and suggestions. Work supported by DARPA, ARD, NSF and ONR.

Correspondence and requests for materials should be addressed to D.D.A. (e-mail: awsch@physics.ucsb.edu).

Jamming phase diagram for attractive particles

V. Trappe*†, V. Prasad*, Luca Cipelletti*†, P. N. Segre*† & D. A. Weitz*

* *Department of Physics and DEAS, Harvard University, Cambridge, Massachusetts 02138, USA*

A wide variety of systems, including granular media, colloidal suspensions and molecular systems, exhibit non-equilibrium transitions from a fluid-like to a solid-like state, characterized solely by the sudden arrest of their dynamics. Crowding or jamming of the constituent particles traps them kinetically, precluding further exploration of the phase space[1]. The disordered fluid-like structure remains essentially unchanged at the transition. The jammed solid can be refluidized by thermalization, through temperature or vibration, or by an applied stress. The generality of the jamming transition led to the proposal[2] of a

† Present addresses: Department of Physics, University of Fribourg, Fribourg, Switzerland (V.T.); Department of Physics, University of Montpellier, Montpellier, France (L.C.); and NASA, Marshall Space Flight Center, Huntsville, Alabama 35812, USA (P.N.S.).

© 2001 Macmillan Magazines Ltd

unifying description, based on a jamming phase diagram. It was further postulated that attractive interactions might have the same effect in jamming the system as a confining pressure, and thus could be incorporated into the generalized description. Here we study experimentally the fluid-to-solid transition of weakly attractive colloidal particles, which undergo markedly similar gelation behaviour with increasing concentration and decreasing thermalization or stress. Our results support the concept of a jamming phase diagram for attractive colloidal particles, providing a unifying link between the glass transition[3], gelation[4,5] and aggregation[6–8].

As an intrinsic parameter to unify the description of all routes to jamming, Liu and Nagel[2] use the density of the system, ρ. For repulsive systems, a non-uniform, applied stress can increase ρ and jam athermal systems[1,9], whereas an osmotic or hydrostatic pressure is required to increase ρ and jam thermal systems; for attractive systems, the interaction itself increases ρ and jams the system. Jamming is overcome and the system is fluidized by an increase in temperature or by a uni-directional stress or load that exceeds the yield stress of the material. Therefore, Liu and Nagel[2] propose ρ^{-1}, temperature, T, and stress, σ, as the axes of a three-dimensional phase diagram, with the jammed state in the inner octant, as shown in Fig. 1.

Here we consider attractive colloidal particles, and treat the suspending fluid as an inert background; thus the density is set explicitly by the particle volume fraction, ϕ. Temperature still thermalizes the system; however, the degree of thermalization is primarily controlled by the interparticle attractive energy, U. In this regard, colloidal particles differ markedly from molecular systems that jam; in colloids, U sets the scale of temperature T, while the density is set independently by ϕ; by contrast, in molecular systems, U and pressure control the density of the system, and T independently determines the thermalization. Thus, for colloidal particles of radius a, the axes of the phase diagram become ϕ^{-1}, k_BT/U and σ/σ_0, where $\sigma_0 = k_BT/a^3$ sets the scale of the stress in weakly attractive colloid systems. Colloidal suspensions allow us to test the concept of a jamming phase diagram as well its applicability to attractive systems.

To emphasize the generality of the concept, we use data from three vastly different colloid systems—carbon black, polymethyl methacrylate (PMMA), and polystyrene. Each system spans a different range of interparticle interactions, and data from all are amalgamated to complete the phase diagram. The essence of the use of a phase diagram to describe the jamming behaviour is summarized by the optical micrographs of thin samples of carbon black shown in Fig. 2. A highly dispersed, fluid phase of particles is transformed into a jammed solid network by increasing ϕ, increasing U, or decreasing σ. In each case, the resultant phase transformation is nearly identical and the structures on the fluid and solid side have very similar appearances. We identify a jammed solid by the existence of a stress-bearing, interconnected network, which results in a low-frequency plateau of the elastic modulus, G'_p (ref. 10).

As shown in Fig. 2a and b for carbon black, a well-defined transition from fluid-like to solid-like behaviour can be found as either ϕ is increased at constant U, or U is increased at constant ϕ. The viscosity, η, diverges as some critical value is approached, whereupon G'_p increases sharply. The phase boundary at ϕ_c can be identified by the critical-like behaviour of both η and G'_p for fixed U (Fig. 2a):

$$\eta = \eta_s(\phi_c - \phi)^{-\nu_\phi} \quad (1)$$

and

$$G'_p = G'_\phi\,(\phi - \phi_c)^{\upsilon_\phi} \quad (2)$$

where $\nu_\phi \approx 0.13$, and $\upsilon_\phi \approx 4.0$ and η_s is the solvent viscosity; both G'_ϕ, and ϕ_c depend on U. Similarly, U_c can be identified by the critical-like behaviour of η and G'_p at fixed ϕ, (Fig. 2b):

$$\eta = \eta_D(U_c - U)^{-\nu_U} \quad (3)$$

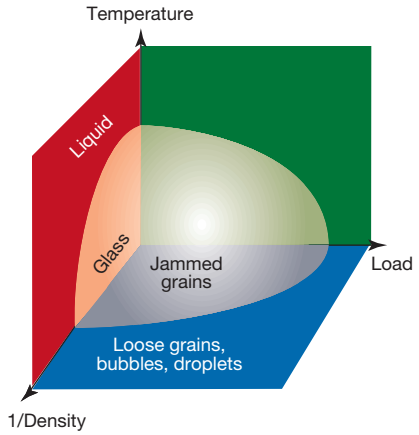


Figure 1 The jamming phase diagram proposed by Liu and Nagel[2].

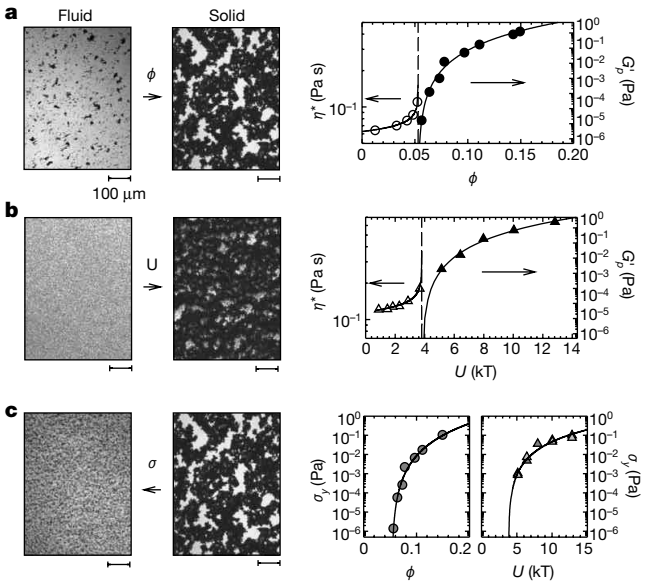


Figure 2 Control parameters for the jamming transition. Optical micrographs showing the jamming phase transition for carbon black going from the fluid-like to solid-like state. **a**–**c**, The volume fraction ϕ increases (**a**), the interaction energy U increases (**b**) and the applied stress σ decreases (**c**). For carbon black there is a divergence of the viscosity and a critical-like onset of the plateau modulus at $\phi_c = 0.053$ for $U \approx 10\,k_BT$ (**a**) and $U_c \approx 3.8k_BT$ for $\phi = 0.14$ (**b**) whereas the yield stress increases critically with both ϕ and U (**c**). The solid lines are results of fits to equations (1), (2), (3), (4), (6) and (7). We note that the points marked as dashed lines in the graphs of **a** and **b** correspond to phase boundaries, while the data themselves are the phase boundaries in the graph of **c**. The exponent for the divergence of the viscosity is rather low; critical-like behaviour, however, is observed over two decades in reduced variables, $\phi_c - \phi$, and $U_c - U$. The values of G'_p are obtained from the scaling behaviour of the frequency-dependent moduli, which allow us to determine very weak values of the plateau modulus, which would otherwise require very low-frequency measurements[10]. The yield stresses are calculated using G'_p and the yield strains, γ_y, which are identified for the onset of strain dependence in $G'(\omega)$ in oscillatory shear measurements. As the phase boundary is approached, γ_y diverges; this behaviour is typical of a gelling system. The carbon black is suspended in a basestock oil, and the interparticle attraction is controlled through the addition of a dispersant, which adsorbs on the surface of the particles, leading to steric stabilization; increasing concentrations of dispersant lead to decreasing values of U (ref. 10). These suspensions are very useful for practical studies, serving as excellent benchtop systems to model the behaviour of soot in lubricating oil.

© 2001 Macmillan Magazines Ltd

and

$$G'_p = G'_U (U - U_c)^{\nu_U} \quad (4)$$

where again $\nu_U \approx 0.13$ and $\upsilon_U \approx 4.0$, and η_D is the viscosity of the fully dispersed system; both U_c and G'_U depend on ϕ. The power-law behaviour is reminiscent of equilibrium critical phenomena; however, while this similarity is intriguing, jamming is clearly a non-equilibrium transition, and we use these functional forms only as convenient descriptions of the behaviour.

Similar scaling behaviour with a critical-like onset of gelation is observed for the ϕ-dependence of PMMA gels, although the value of the exponent υ_ϕ depends on the relative range of the depletion interaction ξ. As the interparticle interaction becomes increasingly attractive, a stress-bearing, interconnected network or gel spanning the system is formed at lower values of ϕ. Dynamic light scattering from these samples shows that the relaxation time due to the motion of clusters of particles diverges as the transition to solid-like behaviour is approached[11]. Thus the solid network is formed by jamming of clusters. The critical values are described by[11]:

$$\phi_c = \phi_0 \exp(-U_c/\alpha k_B T) \quad (5)$$

where α is a coefficient of order unity, and the values of both α and ϕ_0 depend on ξ. We use this functional form to define the boundary in the phase diagram, but note that the behaviour can equally well be described by $\phi \propto U^{-1}$.

In the limit of $U/k_B T \ll 1$, the behaviour approaches that of hard spheres[12], where monodisperse particles jam to form a solid, stress-bearing network at random close packing, $\phi_{rcp} = 0.63$. By contrast, in the limit of $U/k_B T \gg 1$, irreversible aggregation results, leading to the formation of fractal clusters[6]. When the fractal clusters fill space, they jam to form an elastic solid. Because the density of a fractal aggregate decreases as the cluster grows, the system will, in principle, form a gel at any volume fraction[7], so that ϕ_c approaches zero, consistent with equation (5).

The third dimension of the phase diagram is defined by the applied stress, σ, which affects the formation of any solid network spanning the system, shifting ϕ_c to higher values with increasing σ. For example, for the irreversible aggregation of the polystyrene colloids, a gel will not form at low ϕ if the particles are not buoyancy matched; then ϕ_c becomes significantly greater than zero owing to gravitational stress. In general, a sufficiently large stress will cause a jammed solid sample to yield and flow; thus the yield stress, σ_y, defines the phase boundary. We show both the ϕ and the U dependencies of σ_y for the carbon black in Fig. 2c. Remarkably, the data are again well described by critical-like functional forms:

$$\sigma_y = \sigma_\phi (\phi - \phi_c)^{\mu_\phi} \quad (6)$$

and

$$\sigma_y = \sigma_U (U - U_c)^{\mu_U} \quad (7)$$

where $\mu_\phi \approx 3.4$ and $\mu_U \approx 2.4$, whereas the values of the coefficients depend on the values of U (for equation (6)) or ϕ (for equation (7)). These functional forms describe the phase boundaries for the other colloid systems as well, although the values of both the coefficients and exponents depend on the system; we show an example in Fig. 3, where we plot the ϕ-dependence of σ_y for PMMA with different values of U, and for the polystyrene, for which $U/k_B T$ becomes very large, so that $\phi_c = 0$; these data reflect cuts along the phase boundary at different values of U. In all cases, equation (6) describes the functional form, and we use it to define the boundary in the phase diagram.

We plot the full three-dimensional jamming phase diagram for attractive colloidal systems in Fig. 4, where we use the functional forms of equations (5), (6) and (7) to determine the phase boundaries. The values of the coefficients depend on the details of the specific system, and the detailed shape of the phase boundary is likely to be system-dependent. However, the interdependence of U_c, ϕ_c and σ_y is robust, allowing us to plot a schematic phase diagram which is a compendium of contributions from different systems and which correctly captures the overall shape and behaviour. The

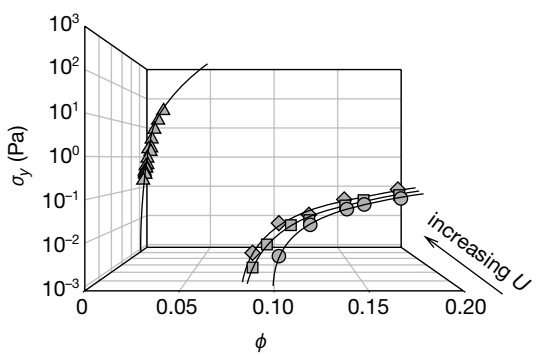


Figure 3 Cuts along the phase boundary defined by σ_y as a function of ϕ for fixed U for polymethyl methacrylate (PMMA) and polystyrene gels. For PMMA, $\xi = 0.28$, $U = 5.0\,k_B T$ (circles), $U = 7.1\,k_B T$ (squares), $U = 8.7\,k_B T$ (diamonds); and for polystyrene gels, $U > 20\,k_B T$. The lines are results of fits to equation (6), with $\mu_\phi \approx 1.5$ for PMMA and $\mu_\phi \approx 3.2$ for polystyrene. All yield stresses are identified from the onset of strain dependence in $G'(\omega)$ in oscillatory shear measurements. The PMMA are stabilized with poly-12-hydroxystearic acid and are suspended in a buoyancy-matched solvent. They are mixed with polystyrene to induce a controlled interparticle attraction owing to the depletion interaction; the strength of the attraction depends on the polymer concentration, whereas its relative range is determined by $\xi = R_g/a$, where R_g is the radius of gyration of the polymer and a is the particle radius[11,14]. The polystyrene latex is also a buoyancy-matched suspension, to which high concentrations of a divalent salt are added to screen the Coulomb repulsion and induce aggregation; this achieves very high values of U (refs 7 and 8).

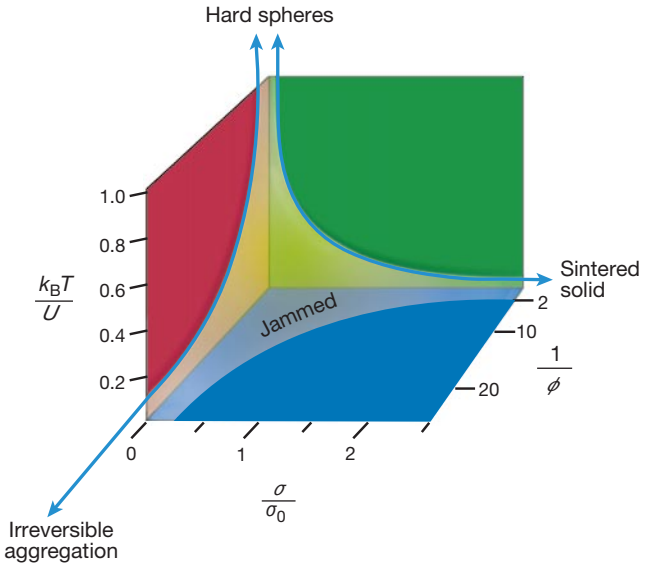


Figure 4 Composite jamming phase diagram for attractive colloidal particles. We focus on $k_B T/U < 1$ and $\phi < 0.5$ and thus do not consider the limits of true hard spheres or of very high concentrations, where more complex behaviour with increasing σ may occur[15]. Data from the three different colloid systems—carbon black, PMMA and polystyrene—are used to construct this phase diagram. It is very difficult to explore the full range of behaviour with any single colloidal system; the phase diagram is therefore a compendium of all systems. We emphasize, however, that the specific phase diagram for any given system may deviate from that shown here, although these deviations will be in the detailed shape, and not the overall behaviour.

© 2001 Macmillan Magazines Ltd

existence of this phase diagram supports the fundamental concept of jamming, and confirms the applicability of jamming to describe the behaviour of attractive particles. However, the shape of the experimental phase diagram differs significantly from that originally proposed by Liu and Nagel[2], having everywhere the opposite curvature, and diverging at each corner. These divergences reflect the particular details of attractive colloidal particles and correspond to irreversible aggregation, where ϕ_c^{-1} is large; to the limit of hard spheres, where T/U is large; and to high volume fractions of strongly attracting particles that form, for example, sintered solids, where σ_y is large.

The basic concept of a jamming phase transition to describe these diverse phenomena is supported by the similarity in the behaviour as the boundary is crossed; variation in ϕ or U leads to equivalent critical-like behaviour for the viscosity (equations (1) and (3)) and the plateau modulus (equations (2) and (4)). Similarly, the yield stress defining the phase boundary exhibits critical-like variation in both ϕ and U (equations (6) and (7)). Moreover, as shown in the micrographs in Fig. 2, the stress plays a similar role to ϕ and U. This suggests that ϕ, U and σ are valid control parameters, and that changes in any of them are equivalent. Additional support for this equivalence comes from the behaviour of the full frequency-dependent storage and loss moduli of solid-like samples, $G'(\omega)$ and $G''(\omega)$, which can be scaled onto identical master curves as either ϕ is varied and U is held fixed, or as U is varied and ϕ is held fixed. Similar scaling is observed for the nonlinear stress response as a function of shear rate for the solid-like samples[10]. Finally, both the divergence of the relaxation time of the clusters as gelation approaches, and the shape of the correlation function, are similar to the behaviour observed as the colloidal glass transition is approached, highlighting the equivalence of gelation and the glass transition as examples of the jamming transition[11].

An integral feature of the jamming transition is the concept of stress-bearing chains[1,9], which was introduced initially for granular systems, but has since been extended to other systems as well[13]. Attractive colloid systems might provide a useful system with which directly to probe these force chains; unlike more dense repulsive systems, the chains are directly visible as the tenuous colloidal network, with nearly all constituents being part of the stress-bearing structure. By contrast, in repulsive systems, force chains are difficult to observe, because they are obscured by the large number of spectator particles, which are not part of the stress-bearing structure. However, gelling systems show a characteristic divergence of the yield strain, γ_y, as the jamming phase boundary is approached, in contrast to granular systems, which are predicted to be fragile, implying a finite γ_y (ref. 9). The differences in behaviour of the two classes of jammed systems merit further investigation.

Our results confirm that a phase diagram can be used to unify the description of a wide variety of transitions from fluid-like to solid-like behaviour by means of the jamming transition, providing striking confirmation of the proposal of Liu and Nagel[2]. Moreover, our results justify the applicability of the jamming transition in describing systems with attractive interactions, significantly extending the range of materials to which these ideas apply. For practical applications, the jamming phase diagram provides important guidance in the parameters that should be changed to achieve a desired behaviour. We conclude by emphasizing that the concept of jamming is a powerful means to account for a wide range of fluid–solid transitions found in colloid systems and helps rationalize apparently diverse behaviour for very different systems; thus, the colloidal glass transition, colloidal gelation and aggregation can all be put on an equal footing as representations of the jamming transition. □

Received 16 January; accepted 30 April 2001.

1. Edwards, S. F. & Grinev, D. V. in *Jamming and Rheology: Constrained Dynamics on Microscopic and Macroscopic Scales* (eds Liu, A. J. & Nagel, S. R.) 80–93 (Taylor and Francis, New York, 2001).
2. Liu, A. J. & Nagel, S. R. Jamming is not just cool anymore. *Nature* **396,** 21–22 (1998).
3. Pusey, P. N. & van Megen, W. Observation of a glass transition in suspensions of spherical colloidal particles. *Phys. Rev. Lett.* **59,** 2083–2086 (1987).
4. Grant, M. C. & Russel, W. B. Volume-fraction dependence of elastic moduli and transition temperatures for colloidal silica gels. *Phys. Rev. E* **47,** 2606–2614 (1993).
5. de Rooij, R., van den Ende, D., Duits, M. H. G. & Mellema, J. Elasticity of weakly aggregating polystyrene latex dispersions. *Phys. Rev. E* **49,** 3038–3049 (1994).
6. Weitz, D. A. & Oliveria, M. Fractal structures formed by kinetic aggregation of aqueous gold colloids. *Phys. Rev. Lett.* **52,** 1432–1435 (1984).
7. Carpineti, M. & Giglio, M. Spinodal-type dynamics in fractal aggregation of colloidal clusters. *Phys. Rev. Lett.* **68,** 3327–3330 (1992).
8. Gisler, T. & Weitz, D. A. Strain hardening of fractal colloidal gels. *Phys. Rev. Lett.* **82,** 1064–1067 (1999).
9. Cates, M. E., Wittmer, J. P., Bouchaud, J. P. & Claudin, P. Jamming, force chains, and fragile matter. *Phys. Rev. Lett.* **81,** 1841–1844 (1998).
10. Trappe, V. & Weitz, D. A. Scaling of the viscoelasticity of weakly attracting particles. *Phys. Rev. Lett.* **85,** 449–452 (2000).
11. Segre, P. N., Prasad, V., Schofield, A. B. & Weitz, D. A. Glass-like kinetic arrest at the colloidal gelation transition. *Phys. Rev. Lett.* (in the press).
12. Pusey, P. N. & van Megen, W. Phase behaviour of concentrated suspensions of nearly hard colloidal spheres. *Nature* **320,** 340–342 (1986).
13. O'Hern, C. J., Langer, S. A., Liu, A. J. & Nagel, S. R. Force distributions near jamming and glass transitions. *Phys. Rev. Lett.* **86,** 111–114 (2001).
14. Pusey, P. N., Pirie, A. D. & Poon, W. C. K. Dynamics of colloid polymer mixtures. *Physica A* **202,** 322–327 (1993).
15. Melrose, J. R. & Ball, R. C. The pathological behaviour of sheared hard spheres with hydrodynamic interactions. *Europhys. Lett.* **32,** 535–540 (1995).

Acknowledgements

We thank S. Nagel and A. Liu for valuable discussions, and for permission to reproduce Fig. 1. This work was supported by Infineum, the NSF and NASA.

Correspondence and requests for materials should be addressed to D.A.W. (e-mail: weitz@deas.harvard.edu).

Formation of chiral morphologies through selective binding of amino acids to calcite surface steps

C.A. Orme*, A. Noy*, A. Wierzbicki†, M. T. McBride*, M. Grantham‡, H.H. Teng§, P.M. Dove‡ & J.J. DeYoreo*

* *Chemistry and Material Science Department, Lawrence Livermore National Laboratory, Livermore, California 94551, USA*
† *Department of Chemistry, University of South Alabama, Mobile, Alabama 36688, USA*
‡ *Department of Geological Sciences, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA*
§ *Department of Earth and Environmental Sciences, The George Washington University, Washington, District of Columbia 20052, USA*

Many living organisms contain biominerals and composites with finely tuned properties, reflecting a remarkable level of control over the nucleation, growth and shape of the constituent crystals[1–6]. Peptides and proteins play an important role in achieving this control[1,7,8]. But the general view that organic molecules affect mineralization through stereochemical recognition, where geometrical and chemical constraints dictate their binding to a mineral, seems difficult to reconcile[4] with a mechanistic understanding, where crystallization is controlled by thermodynamic and kinetic factors[9]. Indeed, traditional crystal growth models emphasize the inhibiting effect of so-called 'modifiers' on surface-step growth, rather than stereochemical matching to newly expressed crystal facets. Here we report *in situ* atomic force microscope observations and molecular modelling studies of calcite growth in the presence of chiral amino acids that reconcile these two seemingly divergent views. We find that enantiomer-specific binding of the amino acids to those surface-step edges that offer the best geometric and chemical fit changes the step-edge

© 2001 Macmillan Magazines Ltd

free energies, which in turn results in macroscopic crystal shape modifications. Our results emphasize that the mechanism underlying crystal modification through organic molecules is best understood by considering both stereochemical recognition and the effects of binding on the interfacial energies of the growing crystal.

Growth of pure calcite, which has been investigated in detail by atomic force microscopy (AFM)[10], yields rhombohedral crystals with six crystallographically equivalent {104} facets. The atomic steps on the crystal surfaces reflect this symmetry (Fig. 1a), with two steps acute to the {104} cleavage plane and two steps obtuse to the cleavage plane, forming a rhombus. The two obtuse steps are related through a glide-plane symmetry element, as are the two acute steps[11].

We monitored the effects of adding different amino acids to the solution from which calcite is grown. As shown in Fig. 1b, upon adding glycine, an achiral amino acid, the two acute steps become curved and lose their well defined facet directions, while the two obtuse steps are unaffected and the growth hillock retains its overall symmetry about the glide plane. Addition of an aqueous aspartic acid (Asp) solution also results in only the two acute steps becoming curved while the two obtuse steps remain unaffected. However, with the addition of only one of the two Asp enantiomers, D- or L-Asp, the symmetry about the glide plane is broken. The surface morphology depends on the chirality of the aspartic acid: growth hillocks formed in the presence of D- and L- forms of Asp are mirror images of one another (Fig. 1c and d). This Asp-induced change in surface morphology manifests itself also in the shape of so-called etch pits, which form when the crystal dissolves (Fig. 1e and f)[12]. As the additive Asp has not been introduced into the bulk crystal, these effects must result from interactions between calcite and Asp at the surface step edges.

Some organic molecules are known to alter the habit of bulk crystals[3,5,6,8,13,14]. To investigate how the changes in the atomic-scale configuration of surface steps illustrated in Fig. 1c–f might translate into bulk crystal habit alterations, we nucleated calcite in the presence of Asp. As illustrated by Fig. 1g and h, a new set of facets is expressed, and these lie approximately parallel to the ⟨001⟩ axis of pure calcite and map to the {hk0} family of planes. Three {104} facets cap the {hk0} cylinder, with the shape of these cap facets matching almost exactly the shape exhibited by the calcite atomic steps seen in Fig. 1c and d. The obtuse steps, unaffected by Asp, form the straight ridges connecting the cap facets, while the curved acute steps form the rounded base of the cap. Changes in the shape of atomic steps can thus directly modify the macroscopic crystal shape of calcite, by inducing the formation of a new family of faces approximately parallel to the *c* axis.

Chiral molecules have been shown to reduce the symmetry of composite systems[12,15–19], and formation of homochiral surface facets upon adsorption of chiral amino acids has recently been

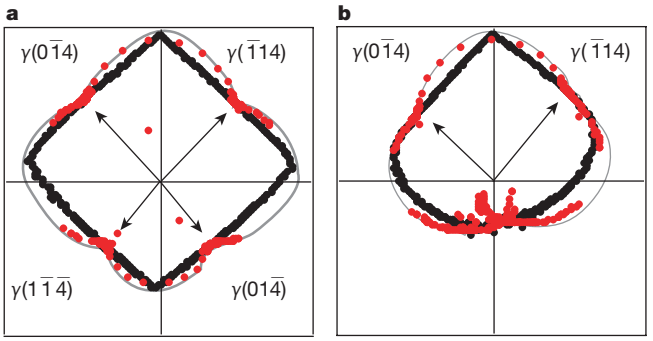


Figure 2 Orientational dependence of step-edge energy (two-dimensional Wulff plot) in the {104} plane. Plots are based on equilibrium crystal shapes for crystals grown in **a**, pure solution and **b**, 0.01 M D-Asp-bearing solution. The step shape is shown in black and the relative step-edge free energy γ calculated from these shapes is shown in red (with a grey line to guide the eye). The binding energy calculations predict that D-Asp will preferentially bind to the (01$\bar{4}$) riser over the (1$\bar{1}\bar{4}$) riser. The new minimum also appears in this quadrant. The two-dimensional Wulff plot for L-Asp-bearing solutions is the mirror image of **b**. The validity of assuming equilibrium crystal shapes in our system is supported by the similarity of the atomic-scale morphology of the growth hillock on the {104} facet to that of the bulk crystal, which suggests that the steps are indeed in their equilibrium geometry. Even hillocks grown in solutions with near-equilibrium $CaCO_3$ concentrations displayed this modified shape, as did the first turn of the spiral which is set by the energetics of the critical nucleus shape.

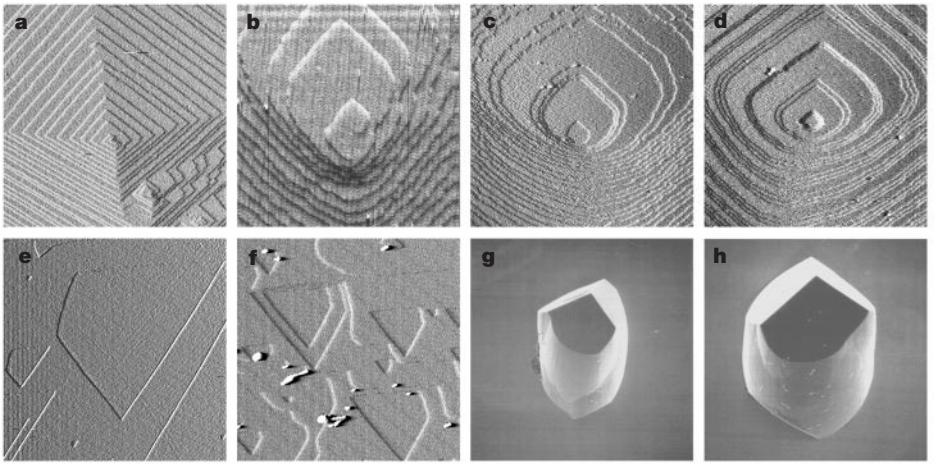


Figure 1 Images showing the effect of amino acids on calcite morphology. **a**–**f**, AFM images showing the effect of amino acids on growth-hillock and dissolution-pit geometry. **a**, A pure calcite growth hillock. **b**–**d**, Growth hillocks following addition of supersaturated solutions with 0.01 M glycine, an achiral amino acid (**b**); 0.01 M L-aspartic acid (**c**); and 0.01 M D-aspartic acid (**d**). **e**,**f**, Dissolution pits undersaturated with respect to calcite and saturated with respect to L- and D-Asp respectively. All images are shown in the same orientation as **a** with two obtuse steps at the top of the image. Image sizes are: **a**, 3.5 × 3.5 μm; **b**, 3 × 3 μm; **c** and **d**, 15 × 15 μm; **e**, 10 × 10 μm; and **f**, 5 × 5 μm. **g**, **h**, SEM micrographs of calcite crystals nucleated on COOH-terminated regions of patterned self-assembled monolayers of alkane thiols[33] and grown in the presence of 0.01 M L-Asp and D-Asp respectively. The macroscopic crystals exhibit a columnar morphology with three {104} facets forming a cap at each end. The straight edges of these caps correspond to the obtuse steps that are the two steps in the upper half of the AFM images. The shape and chirality of each facet matches that of the growth hillocks shown in **c** and **d**. The curved surfaces that form the barrel of each column have general {hk0} orientations.

© 2001 Macmillan Magazines Ltd

reported[20]. Our observations, showing that breaking the symmetry of atomic steps leads to a reduced overall crystal symmetry and the emergence of chiral shapes, elucidates the interactions between chiral molecules and inorganic surfaces that give rise to these effects. Living organisms might use a similar mechanism to create the chiral morphology seen in many biomineralized structures[13,18].

Figure 2 shows polar plots of the free energies of the step edges on the {104} calcite facet (that is, two-dimensional Wulff plots[21]) constructed from the facet shapes (see Fig. 2 legend for details). The Wulff plot for pure calcite growth exhibits distinct minima along the four step directions (Fig. 2a). The hillock growing in the presence of Asp is much more isotropic near the three corners involving the acute steps; in the corresponding Wulff plot, the minima associated with the two acute step edges is eliminated, reflecting that the energetic cost of creating curvature—and therefore kink sites—is significantly diminished in this system. In addition, a new minimum appeared near the intersection of the glide plane with the two acute step edges, indicating that Asp addition generates a new low-energy orientation for the step. The minimum position switched from the right side of the glide plane to the left side upon changing the amino acid from D-Asp to L-Asp.

The addition of Asp at 2 to 200 times the $CaCO_3$ concentration had little or no effect on step-growth speeds, with only extremely high Asp concentrations (~2,000 times the $CaCO_3$ concentration) leading to significant inhibition. Although an altered step geometry (Fig. 1c and d) is eventually obtained, the step edge remains smooth even at these high impurity levels. These observations suggest that in contrast to the effects of impurity species in many other systems (see ref. 23 for example), Asp molecules adsorbed on the calcite surface do not act as pinning sites for step motion that would affect the growth kinetics. For additives that incorporate into the bulk, such as magnesium[23], the step velocity is strongly dependent on the additive concentration. Asp does not show such behaviour and therefore it is unlikely that Asp incorporates into the bulk of calcite crystal. Some kinetic effects may play a role, but the consistency in step shape over time and from the first turn of the spiral to the macroscopic shape, and the results of our step velocities measurements, all argue that growth is altered primarily by changing the energetics at the step edge.

Using X-ray photoelectron spectroscopy (XPS)[24] and surface X-ray diffraction[25–27], no evidence for Asp either on the {104} face or within the bulk of the crystal was observed, and the coverage of Asp on the {104} face and its abundance within the bulk is thus minimal ($\leq$ 0.3%). This finding might be taken to suggest that the effect of Asp on the {104} step morphology is due solely to its interaction with the steps, whose surface density never exceeded ~1%. However, typically only 0.02% protein is occluded in biogenic calcite skeletal elements, and similar occlusion levels have been found in synthetic calcite with altered morphology[28].

The chiral nature of the observed interaction implies that Asp binding to the calcite surface spans the chiral centre of the amino acid. Moreover, if the binding was localized at only one surface site, an identical binding position could be found on the opposite side of the glide plane that, by definition, exhibits mirror symmetry (Fig. 3). Consequently, an individual amino acid molecule must interact with at least two sites on the calcite face to exhibit chiral binding. Finally, lack of evidence for significant amounts of Asp on the {104} faces implies that the interaction only occurs at the step edges, and that it is the structure of these step edges that is modified and results in the expression of near-{hk0} faces.

Similar bidentate substitutional bonding to calcite {hk0} faces has been proposed[5] for molecules having two carboxyl groups (which includes Asp), with binding geometries on the {110} faces that are in general agreement with our results. However, our AFM data clearly show that surface growth occurs on single steps, even in the presence of the amino acids, implying that acidic amino acids do not bind to a single {hk0} crystallographic plane during growth, but rather at the step edges. Furthermore, because two-site binding is necessary to explain chiral effects, the Asp–calcite complex must involve sites on both the step riser and the terrace.

We have calculated the binding energy of L- and D-Asp to the {104} face as well as to the acute step risers of single and double steps in the presence of water[29–32]. The results predict that binding of Asp to the flat {104} surfaces of calcite involves all charged groups of the molecule. As shown in Table 1, there is no energy difference for binding of the D- and L- forms of Asp to the (104) surface of calcite, because both adsorbates can rotate to maximize their binding energy.

The binding geometry of Asp to the acute steps reflects the mirror symmetry of these steps and the chirality of the adsorbate (Fig. 3). (For simplicity we discuss only the binding of D-Asp to the acute steps, with the understanding that the discussion for L-Asp is equivalent but with the opposite acute step.) As shown in Fig. 3, on the cleavage plane, D-Asp binds to the $(01\bar{4})$ riser so that one of

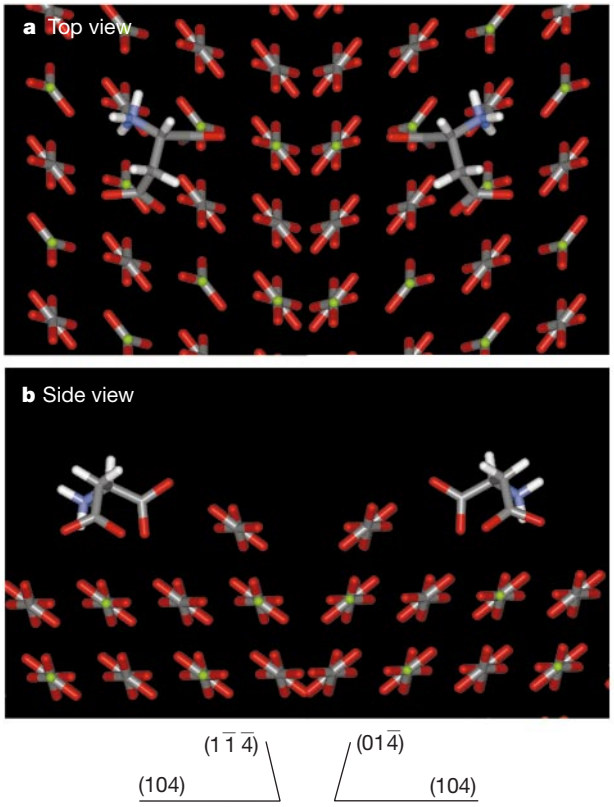


Figure 3 Geometry of binding for aspartic acid adsorbed on the single (104) steps of calcite. **a**, Top view and **b**, side view of L-Asp (left) and D-Asp (right) binding to the steps of calcite with $(1\bar{1}\bar{4})$ risers and $(01\bar{4})$ risers, respectively. The crystal surfaces are rotated and aligned to show the mirror symmetry of binding. Surfaces are viewed along the direction that aligns the calcium ions (small green spheres) with the centres of carbonate ions. Only fragments of the surface used for the modelling are shown. The following atom colouring scheme was used: calcium, green; carbon, grey; oxygen, red; nitrogen, blue; and hydrogen, white.

Table 1 Binding energies of D- and L-aspartic acid to the flat and stepped {104} surfaces of calcite

	(104)	104 single-step $(01\bar{4})$ riser	104 single-step $(1\bar{1}\bar{4})$ riser	104 double-step $(01\bar{4})$ riser	104 double-step $(1\bar{1}\bar{4})$ riser
D-Asp	–34.7	–41.0	–33.1	–49.4	–41.4
L-Asp	–34.3	–32.2	–41.4	–42.3	–49.8

Energies are expressed in units of kJ mol^{-1}. The binding energies were calculated using a dielectric constant of 80.

© 2001 Macmillan Magazines Ltd

its negatively charged carboxyl groups (COO^-) completes the coordination of calcium ions while keeping the positively charged group (NH_3^+) in registry with positive ions at the surface. At the step, the remaining carboxyl group can bind to the two {104} surfaces converging at the step, thus constraining the binding geometry and increasing the binding energy by approximately 19%, relative to that of binding to the cleavage plane. On the $(1\bar{1}4)$ riser, the carboxyl group of D-Asp matched the carbonates of the calcite only poorly; moreover, the group cannot rotate freely to optimize coordination, and the binding energy to the step is therefore slightly lower than to the flat cleavage plane.

When modelling the hypothetical case of Asp binding to double acute steps, we find that D-Asp is again bound more strongly to the $(01\bar{4})$ riser than to the $(1\bar{1}4)$ riser. But a better stereospecific fit to the double step increases the binding energy by approximately 40%, compared to binding to single acute steps on the (104) surface (Table 1). Double steps are not available during growth from dislocation hillocks, but as the steps reach the periphery of each (104) facet, they merge to form the adjacent facet and binding to double steps should become possible. Furthermore, the Asp double-step structure generates pseudo-{hk0} step risers, from which the {hk0}-like surfaces of the bulk crystals are likely to have emerged. These model calculations thus emphasize the energetic basis for the observed morphological changes induced by the presence of Asp during crystal growth.

Our results suggest that the addition of organic growth modifiers alters the energy landscape seen by adsorbing and desorbing calcite species at the surface of the mineral, thus changing the rates of attachment and detachment and/or the surface configuration that is lowest in energy. In particular, we find that site-specific binding of amino acid residues to surface steps changes the step-edge free energies, giving rise to direction-specific binding energies unique to individual amino acid enantiomers and leading to chiral modifications that propagate from atomic length scales to macroscopic length scales. Although stereochemical recognition describes the structural and chemical relationships present during the mineralization process, our findings indicate that the mechanism underlying crystal modification is better understood by considering surfactant-mediated changes to interfacial energies. □

Methods

Solutions and sample preparation

Supersaturated $CaCO_3$ solutions were prepared using reagent grade NaCl, $NaHCO_3$, and $CaCl_2{\cdot}2H_2O$. Saturation ratios (S, defined below) ranged from 1.2 to 3, ionic strength 0.01 M, and pH = 8.1–8.6. Solutions with amino acid additives (10^{-5} M–10^{-2} M) were readjusted for pH. Samples used for XPS and SXRD were prepared by overgrowth on cleaved geologic calcite (Brazil) similar to the AFM experiments. Patterned self-assembled monolayers of mercaptohexadecanoic acid and hexadecanethiol were fabricated on Au-coated (20 Å of Ti, 750 Å of Au) Si substrates using microcontact printing with an elastomeric stamp[33]. The patterned substrates were placed monolayer side down on plastic standoffs into supersaturated calcium carbonate solutions sealed from the atmosphere. This placement ensured that only the crystals that nucleated directly on the film remained on the film. After two days the substrates were taken out of the solution and imaged by SEM. $S = a(Ca^{2+})\,a(CO_3^{2-})/K_{sp}$, where $a(Ca^{2+})$ and $a(CO_3^{2-})$ are the activities of the calcium ion and carbonate ion respectively and K_{sp} is the solubility constant for calcite.

Imaging

Atomic force microscopy (AFM) images were taken in contact mode using a Digital Instruments Nanoscope IIIa AFM equipped with a fluid cell. All images were obtained in solution under flowing conditions. High flow rates (~1 ml min^{-1}) were chosen to ensure surface kinetics were not affected by diffusion limitations. Some step-angle distortion occurred, owing to step-edge migration during the scan time. The images shown are not corrected for this. However, numerical data were corrected for the Doppler shifts by correlating successive images captured in the down and up scan direction. To minimize this effect, typical scan rates were 10 lines per second and thus had a scan time of approximately 50 s per image. Care was taken to ensure that the tip was not modifying the surface during the scan. For example, repeated scanning of the tip over the same point did not cause any appreciable differences at that location relative to the rest of the image. No local erosion or local enhancement of step velocities was observed. At the highest concentrations of Asp tested (0.1 M), lightly bound precipitates formed on the terraces because the Asp was supersaturated; these could be swept away with the tip motion.

X-ray photoelectron spectroscopy (XPS)

X-ray photoelectron spectra were measured using a Perkin-Elmer 5400 ESCA system equipped with a Mg X-ray source (Mg Kα 1253.5 eV). Spectra were recorded using a concentric hemispherical analyser in the on-top position of the chamber, giving an overall energy resolution of 1.1 eV. Survey scans (89.5 eV pass energy, 1 eV step size) and high-resolution XPS spectra (17.90 eV pass energy, 0.1 eV step size) were acquired to determine surface compositions. Calcite and Asp both contain carbon and oxygen; therefore, we used the N 1s photoemission line as a fingerprint to establish the presence or absence of Asp on calcite surfaces. XPS has a detection sensitivity of 0.3 atomic per cent or about 0.5 weight per cent. The kinetic energy of the N 1s core electron is 850 eV, which corresponds to an escape depth of about 10 Å. Thus, the probability of an electron escaping from a N atom, and reaching the detector, from a depth of 10 Å is $1/e$, or about 37%; from a depth of 20 Å, the probability is $(1/e)^2$, or about 13.5%.

Surface X-ray diffraction (SXRD)

X-ray reflectivity measurements were performed at bending-magnet station 1-BM (SRI-CAT) of the Advanced Photon Source at Argonne National Laboratory. The basic principles of the technique have been described elsewhere[25,27]. Crystals were mounted in a custom X-ray reflectivity cell (P. Fenter, M.T.M., G. Srajer, N.C. Sturchio & D. Bosbach, manuscript in preparation), bathed in about 1 ml of mother liquor. Experiments were performed in a static thin-film configuration with a water layer having a typical thickness of about 60 μm at the calcite surface contained by an 8-μm Kapton film. Each data set was measured in less than 2 h. A monochromatic X-ray beam (λ = 0.775 Å; photon energy, 16 keV) was incident upon the surface at a small angle, α (defined with respect to the surface plane), and the incident and reflected photon fluxes were measured separately as a function of α. Typical incident X-ray flux was about 10^{10} photons per second. X-ray reflectivity measurements were performed to a momentum transfer of $Q_{max} \approx 4$ Å^{-1} resulting in a resolution of about 1.5 Å. This provides a simple estimate for the length scale over which the structure can be determined uniquely.

Molecular docking

Monovalent ions of D- and L-Asp were constructed using Spartan software for electronic calculations (Spartan 5.1.1 from Wavefunction Inc., of Irvine, California) and their geometries were optimized using the PM3[29] semi-empirical method. Corrections for aqueous solvation were applied using the Cramer–Truhlar SM3 solvation model[30,34,35]. Electrostatic charges were assigned using the method of fitting to molecular electrostatic potentials[31,32]. Flat and stepped calcite surfaces were built using Cerius[2] (from Molecular Simulations Inc., of San Diego, California) Surface Builder (Version 4.2MS) with the crystal surfaces charges assigned using Cerius[2]. Energy minimizations were performed using the Universal Force Field 1.02 with dielectric constant equal to 80 to account for the water solvation effect. Binding energies were obtained by subtracting from the optimized energy of the adsorbate-surface system the energies of the surface and the adsorbate, when separated beyond the interaction distance.

Received 8 January; accepted 5 April 2001.

1. Mann, S. *et al.* Crystallization at inorganic-organic interfaces—biominerals and biomimetic synthesis. *Science* **261,** 1286–1292 (1993).
2. Addadi, L., Moradian, L. J., Shay, E., Maroudas, N. G. & Weiner, S. A chemical model for the cooperation of sulfates and carboxylates in calcite crystal nucleation—relevance to biomineralization. *Proc. Natl Acad. Sci. USA* **84,** 2732–2736 (1987).
3. Sikes, C. S., Yeung, M. L. & Wheeler, A. P. in *Surface Reactive Peptides And Polymers: Discovery And Commercialization* (eds Sikes, C. S. & Wheeler, A. P.) ACS Symposium Series 444, Ch. 5, 50–71 (ACS Books, Washington DC, 1991).
4. Mann, S. *et al.* Crystallization at inorganic–organic interfaces—biominerals and biomimetic synthesis. *Science* **261,** 1286–1292 (1993).
5. Mann, S., Didymus, J. M., Sanderson, N. P., Heywood, B. R. & Samper, E. J. A. Morphological influence of functionalized and non-functionalized D-alpha,omega-dicarboxylates on calcite crystallization. *J. Chem. Soc. Faraday Trans.* **86,** 1873–1880 (1990).
6. Berman, A., Addadi, L. & Weiner, S. Interaction of sea urchin skeleton macromoloecules with growing calcite crystals—a study of intracrystalline proteins. *Nature* **331,** 546–548 (1988).
7. Berman, A. *et al.* Biological control of crystal texture—a widespread strategy for adapting crystal properties to function. *Science* **259,** 776–779 (1993).
8. Belcher, A. M. *et al.* Control of crystal phase switching and orientation by soluble mollusc-shell proteins. *Nature* **381,** 56–58 (1996).
9. Burton, W. K., Cabrera, N. & Frank, F. C. The growth of crystals and equilibrium structure of their surfaces. *Phil. Trans. R. Soc. Lond. A* **243,** 299–358 (1951).
10. Teng, H. H., Dove, P. M., Orme, C. A. & DeYoreo, J. J. Thermodynamics of calcite growth: Baseline for understanding biomineral formation. *Science* **282,** 724–727 (1998).
11. Reeder, R. J. in *Carbonates: Mineralogy And Chemistry* (eds Reeder, R. J. & Mineralogical Society of America) Vol. II, Ch. 1, 1–47 (Mineralogical Society of America, Washington DC, 1983).
12. Honess, A. P. & Jones, J. R. *Etch figure Investigations with Optically Active Solvents.* Technical Paper 35, 667–721 (Pennsylvania State College, Mineral Industries Experimental Station, 1937).
13. Aizenberg, J. *et al.* Biologically induced reduction in symmetry—a study of crystal texture of calcitic sponge spicules. *Chem. Eur. J.* **1,** 414–422 (1995).
14. Didymus, J. M. *et al.* in *Mechanisms And Phylogeny Of Mineralization In Biological Systems* (eds Suga, S. & Nakahara, H.) 267–271 (Springer, Tokyo/New York, 1991).
15. Weissbuch, I., Addadi, L., Lahav, M. & Leiserowitz, L. Molecular recognition at crystal interfaces. *Science* **253,** 637–645 (1991).
16. Addadi, L. *et al.* A link between macroscopic phenomena and molecular chirality: crystals as probes for indirect assignment of absolute configuration of chiral molecules. *Topics Stereochem.* **16,** 1 (1986).

© 2001 Macmillan Magazines Ltd

17. Madura, J. D. *et al.* Interactions of the D-form and L-form of winter flounder antifreeze peptide with the (201) planes of ice. *J. Am. Chem. Soc.* **116,** 417–418 (1994).
18. Didymus, J. M., Young, J. R. & Mann, S. Construction and morphogenesis of the chiral ultrastructure of coccoliths from the marine alga *Emiliania Huxleyi*. *Proc. R. Soc. Lond. Ser. B* **258,** 237–245 (1994).
19. Mann, S., Sparks, N. H. & Blakemore, R. P. Structure, morphology and crystal growth of anisotropic magnetite crystals in magnetotactic bacteria. *Proc. R. Soc. Lond. B* **231,** 477–487 (1987).
20. Zhao, X. Y., Zhao, R. G. & Yang, W. S. Scanning tunneling microscopy investigation of L-lysine adsorbed on Cu (001). *Langmuir* **16,** 9812–9818 (2000).
21. Wulff, G. Velocity of growth and dissolution of crystal faces. *Z. Kristallogr.* **34,** 449–530 (1901).
22. Land, T. A., Martin, T. L., Potapenko, S., Palmore, G. T. & De Yoreo, J. J. Recovery of surfaces from impurity poisoning during crystal growth. *Nature* **399,** 442–445 (1999).
23. Davis, K. J., Dove, P. M. & De Yoreo, J. J. The role of Mg^{2+} as an impurity in calcite growth. *Science* **290,** 1134–1137 (2000).
24. Stipp, S. L. & Hochella, M. F. Structure and bonding environments at the calcite surface as observed with x-ray photoelectron spectroscopy (XPS) and low energy electron diffraction (LEED). *Geochim. Cosmochim. Acta* **55,** 1723–1736 (1991).
25. Feidenhansl, R. Surface structure determination by X-ray diffraction. *Surf. Sci. Rep.* **10,** 105–188 (1989).
26. Fenter, P. *et al.* Surface speciation of calcite observed in situ by high-resolution X-ray reflectivity. *Geochim. Cosmochim. Acta* **64,** 1221–1228 (2000).
27. Robinson, I. K. X-ray crystallography of surfaces and interfaces. *Acta Crystal. A* **54,** 772–778 (1998).
28. Albeck, S. Aizenberg, J. Addadi, L. & Weiner, S. Interactions of various skeletal intracrystalline components with calcite crystals. *J. Am. Chem. Soc.* **115,** 11691–11697 (1993).
29. Stewart, J. J. P. Optimization of parameters for semiempirical methods. 2. Applications. *J. Comput. Chem.* **10,** 221–264 (1989).
30. Cramer, C. J. & Truhlar, D. G. An scf solvation model for the hydrophobic effect and absolute free energies of aqueous solvation. *Science* **256,** 213–217 (1992).
31. Breneman, C. M. & Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials—the need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **11,** 361–373 (1990).
32. Chirlian, L. E. & Francl, M. M. Atomic charges derived from electrostatic potentials—a detailed study. *J. Comput. Chem.* **8,** 894–905 (1987).
33. Aizenberg, J., Black, A. J. & Whitesides, G. M. Control of crystal nucleation by patterned self-assembled monolayers. *Nature* **398,** 495–498 (1999).
34. Cramer, C. J. & Truhlar, D. G. General parameterized scf model for free energies of solvation in aqueous solution. *J. Am. Chem. Soc.* **113,** 8305–8311 (1991).
35. Cramer, C. J. & Truhlar, D. G. Am1-Sm2 and Pm3-Sm3 parameterized scf solvation models for free energies in aqueous solution. *J. Computer-Aided Molecul. Design* **6,** 629–666 (1992).

Acknowledgements

We thank S. Zepeda, S. Orme and J. Bearinger for assistance with AFM data collection, P. Fenter for performing the SXRD studies, A. Nelson and C. Evans for help with XPS, S. Sikes for helpful discussions and L. Addadi for a careful reading of the manuscript. This work was performed under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory and was supported by a grant to Georgia Institute of Technology from the Division of Geosciences and Engineering, Office of Basic Energy Sciences, Department of Energy.

Correspondence and requests for materials should be addressed to C.O. (e-mail: orme1@llnl.gov).

Evidence for fault weakness and fluid flow within an active low-angle normal fault

J. S. Floyd*, J. C. Mutter*†, A. M. Goodliffe‡ & B. Taylor‡

** Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964; and Department of Earth and Environmental Sciences, Columbia University; and † Columbia Earth Institute, Columbia University, New York 10027, USA*
‡ School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822, USA

Determining the composition and physical properties of shallow-dipping, active normal faults (dips < 35° with respect to the horizontal) is important for understanding how such faults slip under low resolved shear stress and accommodate significant extension of the crust and lithosphere. Seismic reflection images[1] and earthquake source parameters[2] show that a magnitude 6.2 earthquake occurred at about 5 km depth on or close to a normal fault with a dip of 25–30° located ahead of a propagating spreading centre in the Woodlark basin. Here we present results from a genetic algorithm inversion of seismic reflection data, which shows that the fault at 4–5 km depth contains a 33-m-thick layer with seismic velocities of about 4.3 km s^{-1}, which we interpret to be composed of serpentinite fault gouge. Isolated zones exhibit velocities as low as ~1.7 km s^{-1} with high porosities, which we suggest are maintained by high fluid pressures. We propose that hydrothermal fluid flow, possibly driven by a deep magmatic heat source, and high extensional stresses ahead of the ridge tip have created conditions for fault weakness and strain localization on the low-angle normal fault.

We use multichannel seismic (MCS) reflection data acquired aboard the RV *Maurice Ewing* during a 1995 survey and drilling results from Ocean Drilling Program (ODP) Leg 180 (refs 3 and 4; Fig. 1). MCS profiles[1,4] reveal a normal fault that maintains a dip of 25–30° to about 9 km depth and an offset of 10–12 km between the sedimented hanging wall and the Moresby seamount footwall, which is composed of Palaeocene arc-ophiolite gabbro and dolerite[3,5]. Faulting and uplift of Moresby seamount is estimated to have begun within the last 3.5 Myr, based on the first occurrence of metamorphic talus found in the downdropped hanging wall at ODP Site 1108 and an abrupt increase in sedimentation rate that indicates rapid subsidence of the northern margin at this time[3]. This fault is one of the major structures on which continental extension appears to be localized. The close proximity of the fault to the

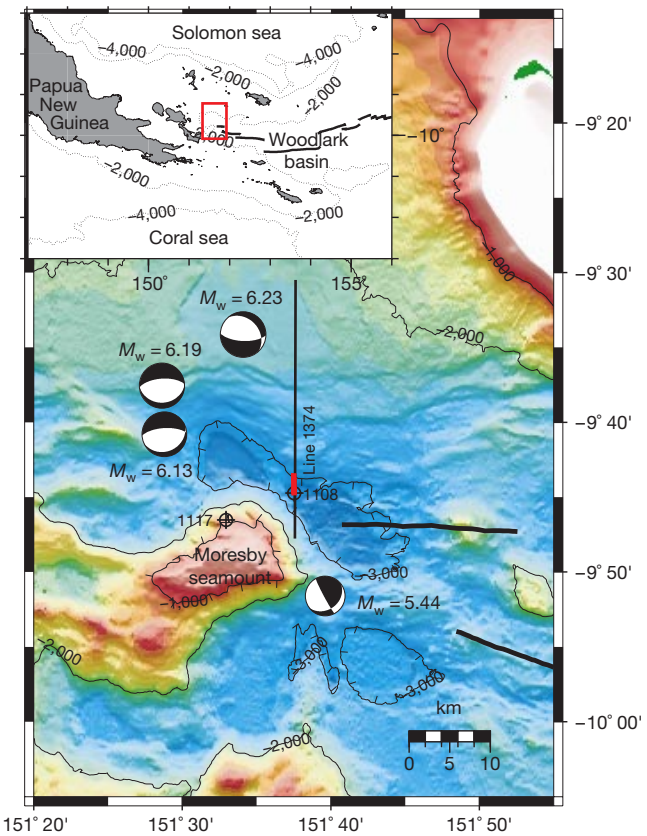


Figure 1 Bathymetry map of the study area showing the location of RV *Maurice Ewing* EW9510 multichannel seismic (MCS) reflection line 1374. The portion of the MCS line 1374 used in the inversion is marked in red. The ODP Leg 180 drill sites 1108 and 1117 discussed in the text are marked[3]. Earthquake focal mechanisms[2] shown include the M_w 6.23 event which occurred on or near the low-angle normal fault on the northern flank of Moresby seamount. Heavy black lines mark the location of the ridge axis interpreted from side-scan sonar data[23]. Bathymetry contours are given in units of metres. Inset, regional map showing the location of the study area (red box).

© 2001 Macmillan Magazines Ltd

spreading ridge tip (Fig. 1) suggests that this will be the last continental fault to form before complete break-up creates new oceanic lithosphere.

Anderson–Byerlee fault mechanics predicts that normal faults should form at dips greater than 45° (ref. 6). Faults may rotate to shallower dips and continue to slip, but for reasonable values of the coefficient of friction[7], they will 'lock up' before reaching 30°. This assumes, however, that the stress field within the fault zone is the same as that outside the fault zone. Rice[8] and others (for example, ref. 9) have shown that a different stress state will exist if the fault contains mechanically weak gouge or pore fluids at near-lithostatic pressures. As seismic velocities are a function of the density, porosity and elastic moduli of the material through which seismic waves propagate, variations in seismic velocity can be used to constrain the composition, physical properties and, hence, the strength of the fault zone.

We investigated the conditions on the imaged fault plane at depth by using a genetic algorithm[10] to invert for the compressional (P) wave velocity and thickness of an approximately 2.15-km-long segment of the fault from the time-migrated MCS line 1374 (Fig. 2). Line 1374 is located on the northeast flank of Moresby seamount and contains the clearest image of the fault compared to other MCS lines in the region which are complicated by out-of-plane reflections from Moresby seamount. The genetic algorithm finds the best-fitting velocity model by allowing a population of starting models to evolve according to the darwinian principle of 'survival of the fittest' (see Methods). We inverted every fifth seismogram spaced 50 m apart between common midpoints (CMPs) 1480 and 1630 between 4.90 and 5.45 s (approximately 4.15 to 5.12 km depth). The theoretical vertical resolution[11] of our MCS data is approximately 12.5 m. The inversion results define a fault zone consisting of a layer 33.4 ± 5.7 m thick with P-wave velocities of 4.3 ± 0.22 km s^{-1} throughout most of the fault and 1.74 ± 0.24 km s^{-1} in isolated sections along the deeper portion of the fault plane (Fig. 3). We also tested a three-layer model to evaluate possible vertical variations in the fault zone velocity. The three-layer model shows a similar velocity structure that contains only minor vertical variations, with the exception of a positive velocity gradient around CMP 1515 and thin 3.0–4.0 km s^{-1} layers interleaved within the low-velocity zones between CMPs 1560 and 1630.

To examine more closely the low-velocity sections of the fault, we analysed the fault reflection amplitude variation with offset (AVO) in CMP gathers 1500 and 1600 (Fig. 4), which are representative of high- and low-velocity sections of the fault zone. At CMP 1500 the P-wave velocity increase at the sediment–fault interface produces a positive polarity reflection at near offsets, whereas at CMP 1600 the P-wave velocity decrease produces a negative polarity reflection. With increasing offset fault reflection amplitudes of both CMP gathers decrease, which corresponds to a decrease in Poisson's ratio across the interface as shown by the Zoeppritz equations[12,13]. Poisson's ratio is a positive, nonlinear function of the ratio between P- and S-wave velocities. The decrease in Poisson's ratio in the low-velocity zones indicates that the P-wave velocity decreases with respect to the S-wave velocity across the interface. This result is best explained by the presence of fluids in the fault zone: because S-waves do not travel through fluids, the S-wave velocity may decrease owing to lower bulk density, but P-wave velocity is reduced by both lower bulk density and slower travel paths through pore fluids. By cross-plotting the zero-offset amplitude (intercept) and rate of change of amplitude with offset (gradient) of the fault zone reflection for CMPs 1480–1630 (ref. 13; Fig. 2), we find that fault reflections with negative intercept and gradient values consistently correlate with low-velocity zones identified by our genetic algorithm inversion (Fig. 3).

The composition of the fault zone material is limited to rock derived from either the sedimentary hanging wall or the gabbroic footwall. The sediment velocity is too low (3 km s^{-1}, Fig. 3) to account for the ~4.3 km s^{-1} fault zone velocity, therefore we infer that the fault zone is composed of material derived from the footwall gabbro that has been reduced in velocity by shearing and hydrothermal alteration. Fluid-assisted deformation of mafic

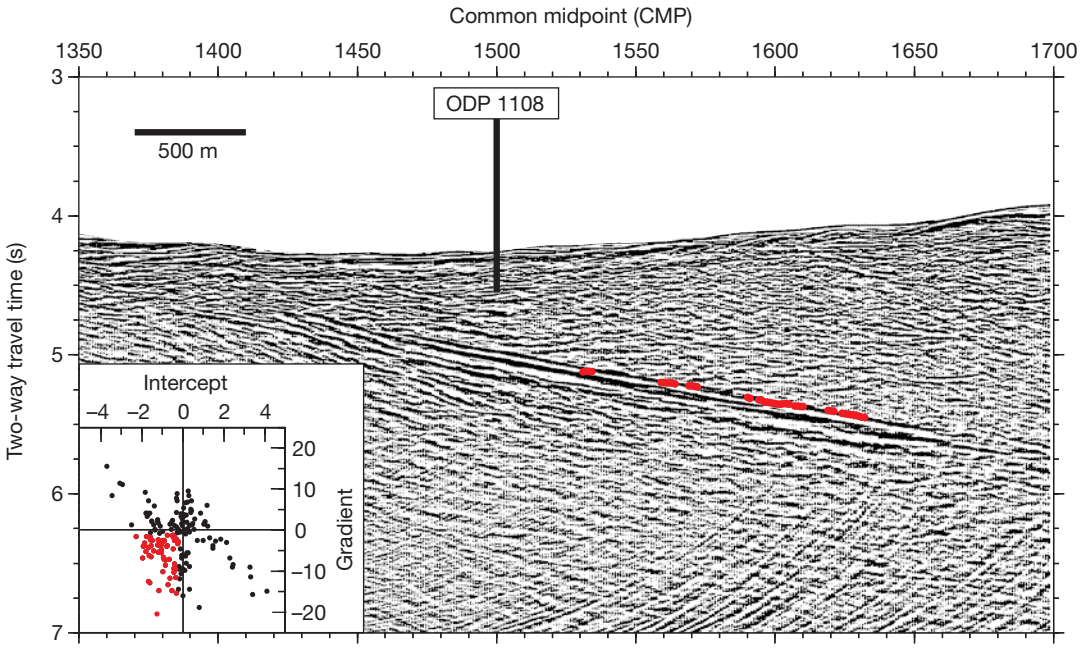


Figure 2 Multichannel seismic line 1374. Processing steps included normal moveout, predictive deconvolution, time migration and spherical spreading correction. ODP Leg 180 site 1108 penetrated to a depth of 485.2 metres below sea floor (m.b.s.f.) and is shown reaching to its approximate depth in units of two-way travel time. CMP spacing is 12.5 m. Inset, intercept and gradient crossplot calculated by fitting the amplitude variation with offset (AVO) values of the fault reflection to the linearized form of the Zoeppritz equations[13] for CMPs 1480 to 1630. Fault reflections with AVO values that fall within the quadrant with negative intercept and negative gradient are marked in red in the crossplot and at the corresponding CMP locations on the seismic profile.

© 2001 Macmillan Magazines Ltd

igneous rocks commonly forms serpentinite, which is widely found in oceanic shear zones along transform and normal faults on mid-ocean ridges[14]. At ODP Site 1117, we recovered approximately 4 m of serpentinized material containing talc, serpentine polymorphs, chlorite and magnetite on top of Moresby seamount, which we interpreted to be fault gouge[3] (Fig. 1). Drilling results confirmed that the northern flank of Moresby seamount is an exposed, ~100-m-thick shear zone that has undergone hydrothermal alteration under greenschist facies metamorphism. We suggest that the fault zone imaged by MCS line 1374 similarly contains serpentinite fault gouge and consider the ~35 m layer determined by seismic waveform inversion to represent only the most deformed and altered material between the offset crustal blocks.

Shipboard laboratory measurements of the serpentinite gouge give a velocity of 2.0 km s^{-1} and a porosity of 30%. Assuming that the low-velocity zones contain material that is similar to the fault gouge recovered at ODP Site 1117, we estimate that a porosity of 61% is needed to explain the decrease in velocity to 1.74 km s^{-1} (ref. 15). This is a minimum value—if higher-velocity material were mixed into the gouge, a higher porosity would be required. Laboratory studies show that only a 10% porosity increase in the fault gouge can be explained by shear deformation[16]. Seismic anisotropy due to shear would lower the overall velocity of the fault zone, but would be expected to be uniform in the direction of slip and would not explain the presence of isolated low-velocity zones. On the basis of our evidence for fluids from genetic algorithm inversion and AVO analysis, we propose that the calculated high-porosity value is best explained by the presence of high pore-fluid pressures in the fault zone. Fluid pressures must be near-lithostatic, otherwise the pore spaces would close under the weight of the overburden.

Evidence for high fluid pressures from our seismic data analysis, together with the occurrence of serpentinite recovered up-dip on the same fault by ODP drilling, suggests that hydrothermal fluid flow may be actively weakening the fault zone. The composition of the fault material and its structural relationship to the nearby spreading ridge lead us to propose that the tectonics of the low-angle normal fault may be explained by analogy to normal detachments flanking inside corner highs of slow- (<25–30 mm yr^{-1}) to intermediate-rate (25–40 mm yr^{-1}) spreading ridges[14,17]. The tip of the Woodlark ridge, which is spreading at a rate of 34 mm yr^{-1} at 151.5° E (ref. 18), is currently located in a right-stepping offset configuration with respect to the normal fault, placing Moresby seamount at a potential inside-corner position. Structural features that are common to inside-corner highs and Moresby seamount include an asymmetric rift axis; a shallowly dipping (20–40°), corrugated[19] slip surface; a rounded dome shape; and a deep nodal basin opposing the elevated footwall[14] (Fig. 1). Larger-scale tectonic elements that are common to the Woodlark rift and oceanic propagators were recognized by Mutter *et al.*[20]. Studies of normal detachments flanking inside corner highs show that they occur at the ends of ridge segments where magma supply is low and

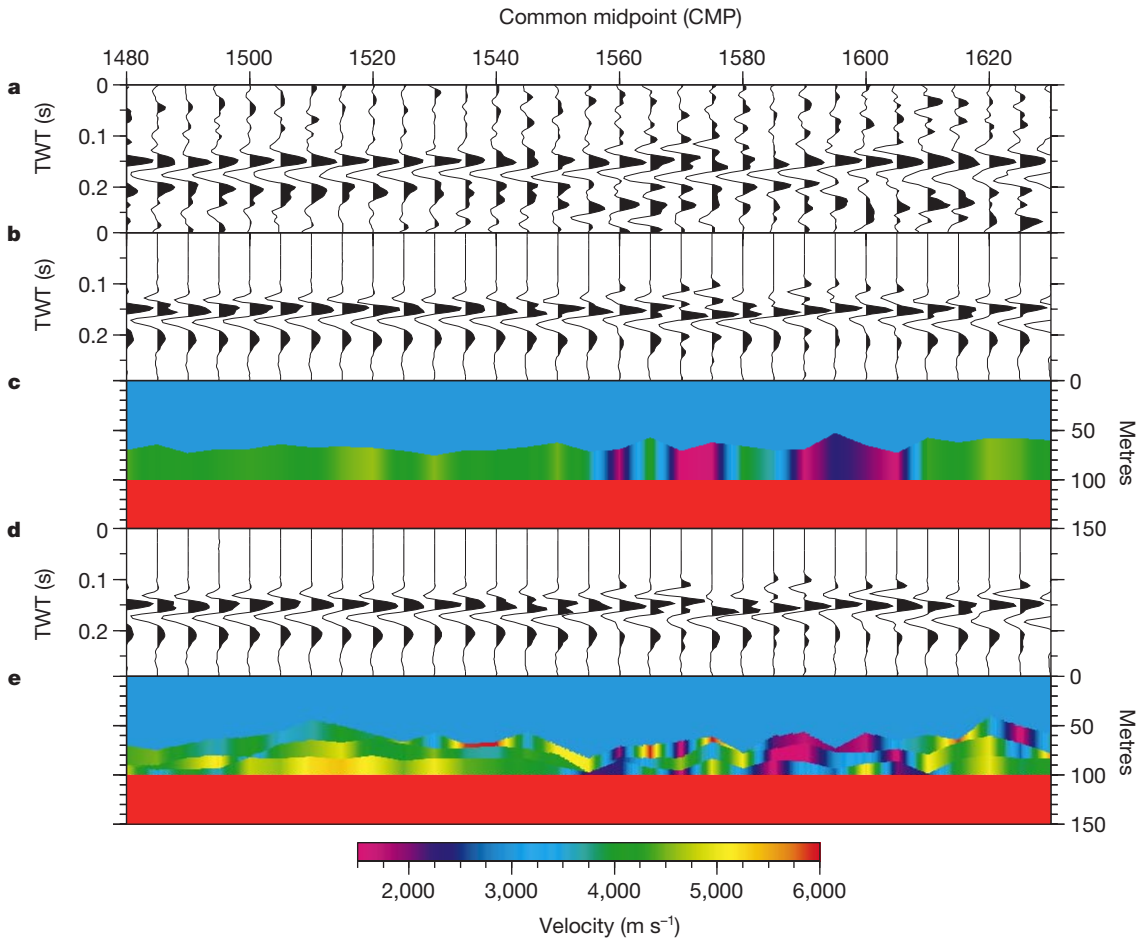


Figure 3 Velocity structure of the fault zone flattened along the fault–basement contact determined by genetic algorithm inversion of every fifth trace from CMPs 1480 to 1630 of MCS line 1374. **a**, Original MCS seismograms flattened at the first positive-polarity peak at or slightly above the fault reflection. **b**, Best-fit synthetic seismograms for the single-layer velocity model (**c**). **d**, Best-fit synthetic seismograms for the three-layer velocity model (**e**). Constant velocities of 3.0 km s^{-1} and 6.0 km s^{-1} were used in the inversion for the sedimentary hanging wall and gabbroic footwall, respectively. TWT, two-way travel time.

© 2001 Macmillan Magazines Ltd

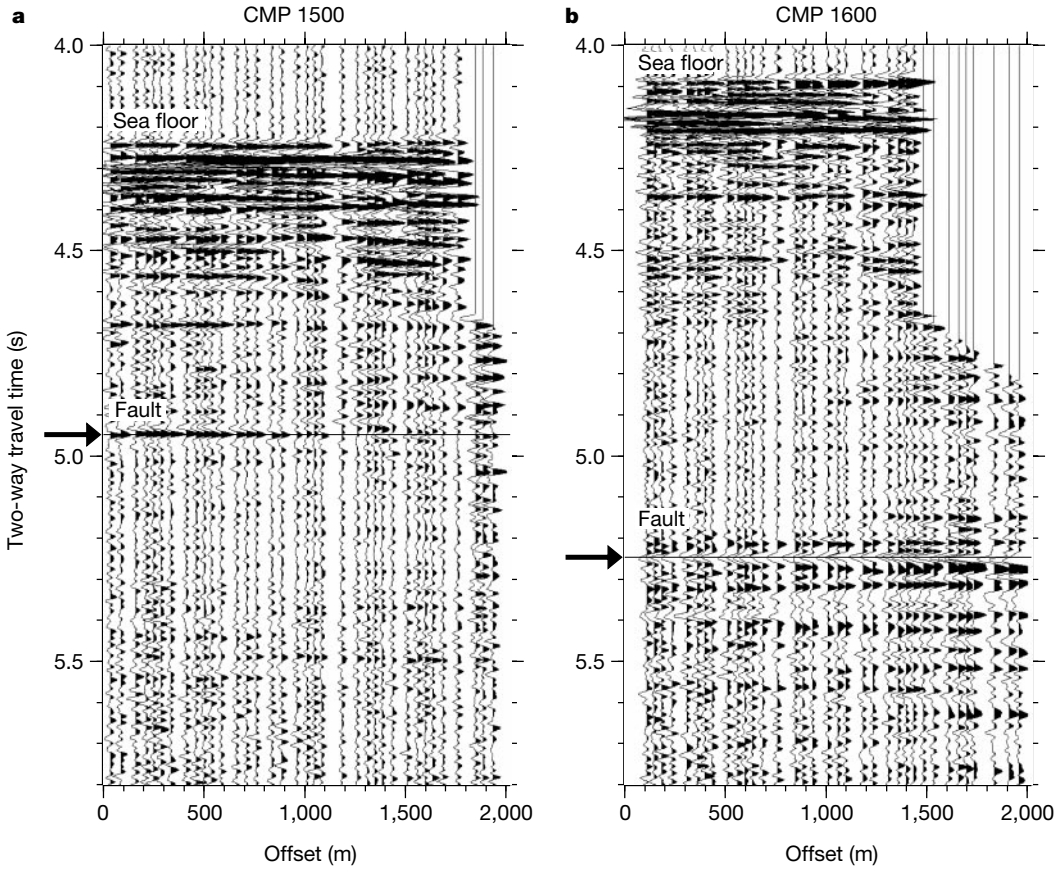


Figure 4 Normal-moveout-corrected CMP gathers 1500 (**a**) and 1600 (**b**) from MCS line 1374. Arrows mark the location of the positive-polarity fault reflection at CMP 1500 and the negative-polarity reflection at CMP 1600. The fault reflection in CMP 1500 (**a**) decreases with offset and exhibits a polarity reversal at 1.2 km, while the fault reflection in CMP 1600 (**b**) that images the low-velocity zone becomes increasingly negative for all offsets.

spreading is accommodated by tectonic extension[14]. ODP Leg 180 borehole temperature measurements at Site 1108 revealed a thermal gradient up to 100 °C km^{-1} in the hanging wall ahead of the ridge tip[3]. We suggest that the role of the Woodlark ridge is to provide high extensional stress and a magmatic heat source that drives convective hydrothermal fluid flow through the fault zone, which weakens the fault by hydrothermal alteration and development of high fluid pressures. In this way, the composition and physical properties of the fault zone and the processes by which these elements form not only provide conditions favourable for normal faulting at low angles, but may also fundamentally control the localization of strain on low-angle normal detachments at the rifting–spreading transition. □

Methods

Genetic algorithms are used in many areas of science to solve nonlinear inverse problems[21]. The genetic algorithm begins with a random population of starting models that undergo a process of selection, crossover and mutation in which the best-fitting models stand the greatest likelihood of being reproduced and passed on to the next generation. The efficiency of the algorithm lies in the fact that unpromising areas of the model space are quickly left behind in favour of more promising areas that are explored in later generations. The process is stopped when the population converges to a maximum fitness value and does not improve with more iterations.

We used the root-mean-square (r.m.s.) error between the synthetic and recorded seismograms as the fitness function to be minimized by the genetic algorithm. Normal incidence synthetic seismogram calculations included the effects of interbed multiples, seismic attenuation and dispersion[22]. The quality factor (Q) of the sediments overlying the fault zone was estimated from MCS data using the spectral ratio method and was found to follow an expected exponential trend with depth from 20 to 65. ODP Leg 180 shipboard measurements[3] and MCS interval velocities provided constraints on the sediment and gabbro velocities, which were held constant in the inversion at 3.0 and 6.0 km s^{-1}, respectively. Starting models were chosen randomly from velocity and thickness combinations between 1.5 km s^{-1} (the velocity of water) to 6.0 km s^{-1} (the velocity of undeformed gabbro) and 0 to 50 m, respectively. Initial tests showed that the best-fit layer thickness was less than 50 m; therefore, the upper layer thickness boundary was set to this value to speed convergence. The starting population for the single-layer model consisted of 100 randomly generated models that evolved over 50 generations to test a total of 5,000 models for each CMP. For the three-layer model, the initial population consisted of 100 individuals that evolved over 100 generations to test 10,000 models for each CMP. The use of one-dimensional seismograms to model a two- or even three-dimensional structure is clearly a simplification; however, this approach is justified by the shallow dip and planar structure of the fault within the modelled region and provides a computational advantage that allowed many more models to be tested. AVO analysis of the CMP gathers provided an independent approach to determination of fault properties that were found to be consistent with the genetic algorithm results and verified that the low-velocity zones are not artefacts of either MCS data processing or the genetic algorithm inversion method.

Received 15 December 2000; accepted 19 April 2001.

1. Mutter, J. C., Mutter, C. Z., Abers, G. & Fang, J. Seismic images of low-angle normal faults in the western Woodlark Basin where continental extension yields to seafloor spreading. *Eos Trans AGU* **74,** 412 (1993).
2. Abers, G. A., Mutter, C. Z. & Fang, J. Shallow dips of normal faults during rapid extension: Earthquakes and normal faults in the Woodlark-D'Entrecasteaux rift system, Papua New Guinea. *J. Geophys. Res.* **102,** 15301–15317 (1997).
3. Taylor, B. *et al. Proc. ODP Init. Rep.* [CD-ROM] **180**, (Ocean Drilling Program, College Station, 1999).
4. Taylor, B., Mutter, C., Goodliffe, A. & Fang, J. Active continental extension: the Woodlark Basin. *JOI/USSAC Newsl.* **9,** 1–4 (1996).
5. Davies, H. L. & Smith, I. E. Geology of Eastern Papua. *Geol. Soc. Am. Bull.* **82,** 3299–3312 (1971).
6. Anderson, E. M. *The Dynamics of Faulting and Dyke Formation with Application to Britain* (Oliver and Boyd, Edinburgh, 1942).
7. Byerlee, J. D. Friction of Rocks. *Pure Appl. Geophys.* **116,** 615–626 (1978).
8. Rice, J. R. in *Fault Mechanics and Transport Properties of Rock* (eds Evans, B. & Wong, T.-F.) 475–503 (Academic, San Diego, 1992).
9. Byerlee, J. D. Friction, overpressure and fault normal compression. *Geophys. Res. Lett.* **17,** 2109–2112 (1990).
10. Sen, M. K. & Stoffa, P. L. Rapid sampling of model space using genetic algorithms: examples from seismic waveform inversion. *Geophys. J. Int.* **108,** 281–292 (1992).
11. Widess, M. B. How thin is a thin bed? *Geophysics* **49,** 1637–1648 (1984).
12. Ostrander, W. J. Plane-wave reflection coefficients for gas sands at nonnormal angles of incidence. *Geophysics* **49,** 1637–1648 (1984).

© 2001 Macmillan Magazines Ltd

13. Shuey, R. T. A simplification of the Zoeppritz equations. *Geophysics* **50,** 609–614 (1985).
14. Tucholke, B. E. & Lin, J. A geological model for the structure of ridge segments in slow spreading ocean crust. *J. Geophys. Res.* **99,** 11937–11958 (1994).
15. Wyllie, M. R. J., Gregory, A. R. & Gardner, G. H. F. An experimental investigation of factors affecting elastic wave velocities in porous media. *Geophysics* **23,** 459–493 (1958).
16. Marone, C., Raleigh, C. B. & Scholz, C. H. Frictional behavior and constitutive modeling of simulated fault gouge. *J. Geophys. Res.* **95,** 7007–7025 (1996).
17. Mutter, J. C. & Karson, J. A. Structural processes at slow-spreading ridges. *Science* **257,** 627–634 (1992).
18. Taylor, B., Goodliffe, A. M. & Martinez, F. How continents break up: Insights from Papua New Guinea. *J. Geophys. Res.* **104,** 7497–7512 (1999).
19. Tucholke, B. E., Lin, J. & Kleinrock, M. C. Megamullions and mullion structure defining oceanic metamorphic core complexes on the Mid-Atlantic Ridge. *J. Geophys. Res.* **103,** 9857–9866 (1998).
20. Mutter, J. C., Mutter, C. Z. & Fang, J. Analogies to oceanic behaviour in the continental breakup of the western Woodlark Basin. *Nature* **380,** 333–336 (1996).
21. Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989).
22. Ganley, D. C. A method for calculating synthetic seismograms which includes the effects of absorption and dispersion. *Geophysics* **46,** 1100–1107 (1981).
23. Taylor, B., Goodliffe, A., Martinez, F. & Hey, R. Continental rifting and initial sea-floor spreading in the Woodlark basin. *Nature* **374,** 534–537 (1995).

Acknowledgements

We thank W. Menke, C. Scholz, S. Carbotte, A. Lerner-Lam and D. Goldberg for helpful discussions and R. Detrick and W. Bosworth for comments that improved this Letter. J.F. and J.M. were supported under a grant from JOI/USSAC. We thank the captain and crew of the RV *Maurice Ewing* for MCS data acquired on EW9510. ODP Leg 180 drilling results used in this study were obtained thanks to the efforts of the captain and crew of the DV *JOIDES Resolution*; A. Klaus and the TAMU scientific support staff; co-chiefs P. Houchon and B. Taylor and the ODP Leg 180 Scientific Party.

Correspondence and requests for materials should be addressed to J.F. (e-mail: jsfloyd@ldeo.columbia.edu).

Anisotropy of thermal diffusivity in the upper mantle

Andréa Tommasi*, Benoît Gibert*, Ulfert Seipold† & David Mainprice*

* *Laboratoire de Tectonophysique, CNRS/Université de Montpellier II, F-34095 Montpellier cedex 5, France*
† *GeoForschungsZentrum Potsdam, Telegrafenberg, D-14473 Potsdam, Germany*

Heat transfer in the mantle is a key process controlling the Earth's dynamics. Upper-mantle mineral phases, especially olivine, have been shown to display highly anisotropic thermal diffusivity at ambient conditions[1,2], and seismic anisotropy data[3] show that preferred orientations of olivine induced by deformation[4] are coherent at large scales (>50 km) in the upper mantle. Thus heat transport in the upper mantle should be anisotropic. But the thermal anisotropy of mantle minerals at high temperature[1,5] and its relationship with deformation have not been well constrained. Here we present petrophysical modelling and laboratory measurements of thermal diffusivity in deformed mantle rocks between temperatures of 290 and 1,250 K that demonstrate that deformation may induce a significant anisotropy of thermal diffusivity in the uppermost mantle. We found that heat transport parallel to the flow direction is up to 30 per cent faster than that normal to the flow plane. Such a strain-induced thermal anisotropy implies that the upper-mantle temperature distribution, rheology and, consequently, its dynamics, will depend on deformation history. In oceans, resistive drag flow would result in lower vertical diffusivities in both the lithosphere and asthenosphere[6] and hence in less effective heat transfer from the convective mantle. In continents, olivine orientations frozen in the lithosphere may induce anisotropic heating above mantle plumes, favouring the reactivation of pre-existing structures.

To investigate the interdependence between flow and thermal anisotropy, that is, to introduce this anisotropy in geodynamical models and quantify its effects, we need to be able to predict the anisotropy of thermal diffusivity that will result from a given upper-mantle flow. Polycrystal plasticity models allow the simulation of the development of olivine lattice-preferred orientations in the upper mantle[6–9]. If the crystallographic orientations and the single-crystal tensor of an anisotropic physical property are known, we may estimate the three-dimensional distribution of this property in a polycrystalline aggregate by averaging the individual grain tensors[10]. Indeed, a strong correlation between olivine fabric and thermal diffusivity anisotropy was observed in a dunite at ambient conditions[1]. Temperature derivatives of the olivine thermal diffusivity tensor are nevertheless poorly constrained. Whereas thermal diffusivity data for olivine and pyroxene crystals showed a constant anisotropy independent of temperature[1], more recent thermal conductivity determinations in olivine crystals[5] exhibited a decrease in anisotropy with increasing temperature, similar to that observed for quartz[11]. Moreover, this simple petrophysical model does not consider parameters that may influence heat transport at the sample and larger scales such as a variation in thermal diffusivity at grain boundaries. These effects as well as the relationship between crystallographic orientation and anisotropy may nevertheless be constrained by comparison between model predictions and thermal diffusivity measurements on real rock samples.

We have thus determined the three-dimensional thermal diffusivity of three mantle rocks by both petrophysical modelling and laboratory measurements between 290 and 1,250 K. Two spinel lherzolites (BALM4 and BALD1) from the Balmuccia and Baldissero massifs in the Italian Alps were chosen as representative

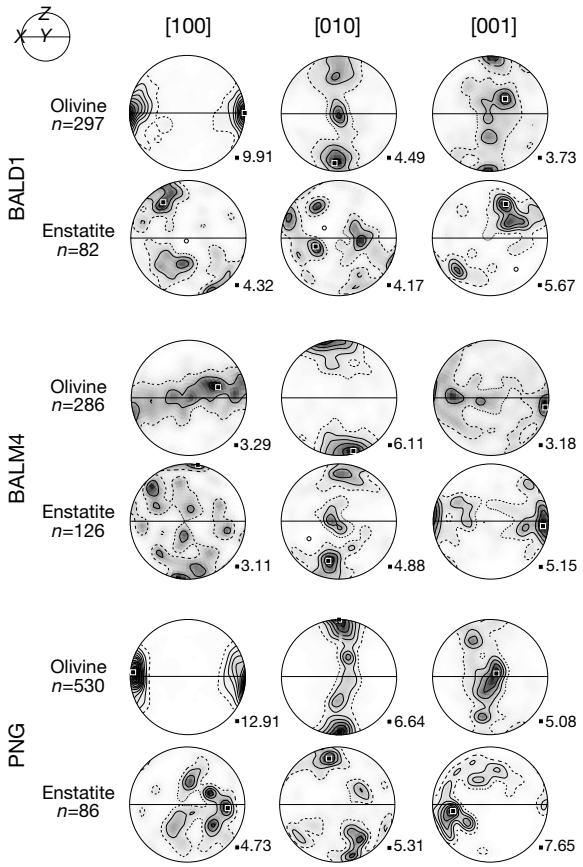


Figure 1 Olivine and enstatite lattice-preferred orientations for the studied samples. Lower hemisphere equal-area projection, *n* measurements, contour intervals are at multiples of a uniform distribution. Inverse-log shading varies from white (minimum density) to black (maximum density, also indicated by a solid square). Full line marks the foliation (*X*–*Y* plane); lineation (*X* direction) is horizontal.

© 2001 Macmillan Magazines Ltd

of the subcontinental mantle and a spinel harzburgite (PNG) from the Papua New Guinea ophiolite was selected as representative of the suboceanic mantle. Modal compositions are: BALM4 (76% olivine, 19% enstatite, 3% diopside and 2% spinel), BALD1 (74% olivine, 18% enstatite, 6% diopside and 2% spinel), and PNG (87% olivine, 12% enstatite and 1% spinel). All samples display coarser-grained porphyroclastic textures characteristic of deformation under high-temperature conditions (>1,100 °C) and are devoid of alteration minerals and macroscopic fractures. Olivine displays a bimodal grain size distribution: porphyroclasts range from 5 to 2 mm, whereas recrystallized grains are smaller than 1 mm. Pyroxenes generally form clusters that may reach up to 5 mm. Foliation and lineation are defined by alignment of spinel crystals accompanied, in the lherzolites, by a weak shape-preferred orientation of olivine. In PNG, recrystallization by subgrain rotation leads to elongation of olivine porphyroclasts normal to the spinel lineation.

Lattice-preferred orientations (Fig. 1) are typical of peridotites deformed under high-temperature conditions, in which (010)[100] and (001)[100] are the dominant slip systems for olivine[9]. BALD1 and PNG display a strong concentration of olivine [100] axes subparallel to the lineation (flow direction) and a distribution of both [010] and [001] axes in a plane normal to the lineation, with secondary concentrations normal and parallel to the foliation (flow plane), respectively. BALM4, on the other hand, is characterized by a planar distribution of [100] and [001] along the foliation and a [010] concentration normal to it. Enstatite preferred orientations are weaker. BALM4 and PNG are characterized by a concentration of the main glide direction [001] close to the lineation and a weak concentration of [010] at high angle to the foliation. BALD1 enstatite preferred orientation is oblique to the foliation. Diopside displays very weak preferred orientations.

Petrophysical modelling using olivine and pyroxene thermal diffusivity tensors determined by spectroscopic methods at ambient conditions[2] indicates that all samples display a significant anisotropy of thermal diffusivity (between 17.5 and 28.6%; Fig. 2). Thermal diffusivity is maximum parallel to the olivine [100] concentration, that is, parallel to the lineation (X direction), and minimum parallel to the [010] concentration, that is, normal to the foliation (Z direction). Anisotropy depends on the strength of olivine lattice-preferred orientation; PNG displays the strongest lattice-preferred orientation and the most anisotropic behaviour. Thermal anisotropy also depends on the pyroxene content, which tends to weaken the anistropy; this effect is enhanced when olivine and estatite lattice-preferred orientations are uncorrelated, as in BALD1. In addition, thermal diffusivity of olivine varies more between [010] and [001] directions than between [001] and [100]. Thus the dispersion of [010] and [001] in a plane normal to the lineation (Y–Z plane) showed by BALD1 significantly weakens its anisotropy, whereas the dispersion of [100] and [001] in the foliation (X–Y plane) displayed by BALM4 only slightly reduces its maximum thermal diffusivity. Similar patterns, but higher anisotropies (between 23.9 and 38.4%), are predicted if more anisotropic olivine and enstatite diffusivity tensors[1] are used.

Modelling suggests that, at ambient conditions, upper-mantle rocks retain one-third to half of the olivine single crystal anisotropy. Is this anisotropy preserved at upper-mantle conditions? Does dispersion at grain boundaries affect the thermal diffusivity tensor? In order to answer these questions, we measured the thermal diffusivity of our samples at ambient pressure and temperatures between 290 and 1,250 K. Six measurements were performed parallel to the lineation (X direction, [100] concentration, for which the highest diffusivity is predicted), four normal to the flow plane (Z direction, [010] concentration), and two in the flow plane, but normal to the lineation (Y direction, [001] concentration). These measurements (Fig. 3) show a significant anisotropy of thermal diffusivity that remains roughly constant over the entire temperature range, in agreement with high-temperature thermal diffusivity data on olivine and pyroxene crystals[1]. Measured anisotropies (21% for BALM4 and 12% for BALD1) are consistent with predictions of petrophysical models using data from ref. 2 (Fig. 2). The higher anisotropy (31%) previously observed in a dunite[1] may be explained by the lower pyroxene content and very strong crystallographic orientation[12] of that sample (similar to PNG, for which 28% of anisotropy was predicted).

Maximum and minimum thermal diffusivities are coherent from sample to sample and their relative magnitudes match model predictions very well: $K_{\mathrm{PNG\text{-}X}} > K_{\mathrm{BALD1\text{-}X}} \approx K_{\mathrm{BALM4\text{-}X}} > K_{\mathrm{BALD1\text{-}Y}} > K_{\mathrm{BALM4\text{-}Z}}$. Measured thermal diffusivities at 290 K are, however, systematically lower by 20% than model predictions; this difference may be attributed to isotropically distributed lattice imperfections

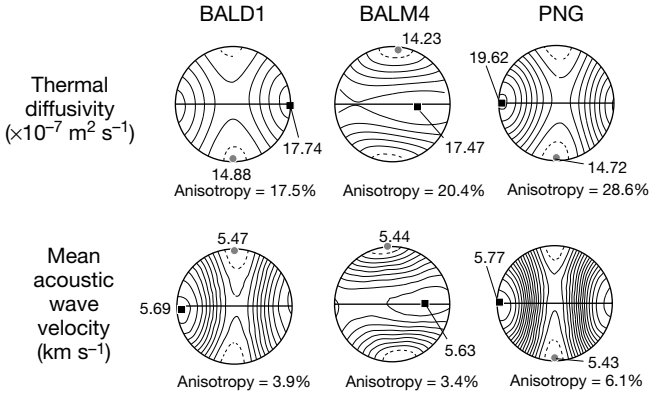


Figure 2 Modelled three-dimensional thermal diffusivity and mean acoustic wave velocity distributions. Voigt–Reuss–Hill averages calculated from crystallographic orientation data and thermal diffusivity and elastic constants tensors for olivine and enstatite at ambient conditions[2,27,28]. As its full thermal diffusivity tensor is not known, diopside is assimilated to enstatite. $V_{\mathrm{mean}} = 3/(1/V_{\mathrm{P}} + 1/V_{\mathrm{S1}} + 1/V_{\mathrm{S2}})$. Lower hemisphere equal-area projections, contours at 10^{-7} m^2 s^{-1} and 0.02 km s^{-1} intervals, respectively. Full line marks the foliation (X–Y plane); lineation (X direction) is horizontal.

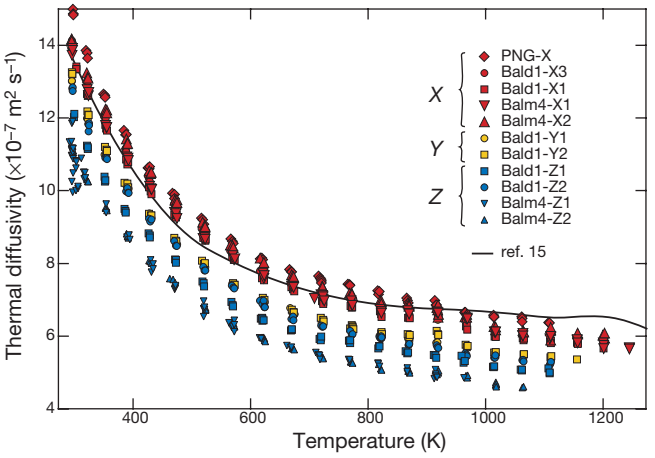


Figure 3 Thermal diffusivity as a function of temperature as measured in oriented cores of PNG (diamonds), BALD1 (circles and squares), and BALM4 (triangles). Red symbols mark measurements parallel to the [100] axis maximum (X direction = lineation), blue symbols indicate measurements parallel to the [010] axis maximum (Z direction = normal to the foliation), and yellow symbols show measurements in the Y direction for BALD1. At each temperature, three successive measurements are performed in order to check data reproducibility. High dispersion in BALM4-Z1 low-temperature data results from electronic noise, eliminated in subsequent measurements. Error bars for individual measurements (not plotted) are of 2%. Relatively high thermal diffusivities displayed by BALD1 in the Z direction result from the dispersion of [010] and [001] in the Y–Z plane (compare Fig. 1). For comparison, the full black line represents thermal diffusivity data for a fine-grained isotropic olivine polycrystal at 100 MPa[15].

© 2001 Macmillan Magazines Ltd

and microfractures in the samples. Indeed, thermal diffusivity measurements in crystalline rocks under increasing pressure conditions generally show a fast, non-linear increase in thermal diffusivity up to 200 MPa associated with closure of microfractures[13,14]. Thus petrophysical models using olivine and pyroxene thermal diffusivity tensors from ref. 2, our experimental constraints on the temperature-derivatives of the olivine and pyroxene thermal diffusivity tensors, as well as a weak linear dependence on pressure[2,15], should provide good estimates of the upper-mantle thermal diffusivity.

Experiments were carefully carried out to minimize problems with radiative disequilibrium. Samples are large (heater–thermocouple distance is 6–7 mm) and low pulse amplitudes ($<$1.5 K at the thermocouple location) were used. Although the measurement distance is close to the photon mean free path in olivine crystals ($\sim$5 mm at 800 K; refs 16 and 17), in real mantle rocks, the latter should be reduced by scattering at grain boundaries and presence of more absorbing phases, like pyroxenes[16,17]. Thus an equilibrium diffusion process should control radiative heat transfer in the experiments and measured thermal diffusivity represents the sum of phonon and photon contributions. This hypothesis is supported by agreement between our data and thermal diffusivity measurements on olivine-rich polycrystalline samples with smaller grain sizes and dimensions[15,18].

Relative contributions of phonon conduction and radiative transfer to the total diffusivity may be inferred from the temperature dependence of thermal diffusivity. Up to 600 K, thermal diffusivity is inversely proportional to temperature, suggesting that scattering of heat-conducting phonons by lattice imperfections and phonon–phonon interactions controls heat transport[19]. At higher temperature, deviation of this behaviour may be ascribed to the onset of radiative heat transfer. Above 900 K, this contribution (about 15% of the total diffusivity) is restrained by development of oxide films along grain boundaries (observed optically in thin sections of the measured cores). Thus lattice conduction is the dominant heat transfer mechanism over the entire temperature range of our measurements. Lattice conduction should also prevail in the Earth's upper mantle, as suggested by analysis of infrared absorption data[20]. Yet, although it is minor, the radiative contribution at high temperature may have been underestimated in our experiments.

Heat transfer by phonons depends essentially on phonon velocity and mean free path[21]. Anisotropy of phonon velocities, which may be approximated by the anisotropy of acoustic velocities[1], is significantly lower than the anisotropy of thermal diffusivity (Fig. 2). Thus the observed thermal diffusivity anisotropy largely results from anisotropy of the phonon mean free path. This latter is controlled by interactions with lattice imperfections and phonon–phonon collisions. At high temperature, however, the contribution of phonon–phonon interactions is dominant. Preservation of a constant anisotropy over the entire temperature range suggests therefore that phonon mean free path and, hence, anisotropy are controlled by phonon–phonon collisions, which are more frequent along the [010] direction. Indeed, linear thermal expansion coefficients for olivine, which are also related to the anharmonicity of lattice potential, display a similar anisotropy, $\alpha_{[100]} < \alpha_{[001]} < \alpha_{[010]}$ up to 2,000 K (ref. 22).

Through the association of petrophysical modelling and thermal diffusivity measurements on textured upper-mantle rocks, we show that deformation may induce a significant thermal diffusivity anisotropy in the uppermost mantle: heat transport parallel to the olivine [100] axes concentration (flow direction) is up to 30% faster than normal to the flow plane ([010] concentration). Moreover, because the anisotropy does not depend on temperature, it might be preserved under higher-temperature, asthenospheric conditions. Existence of a large-scale, strain-induced thermal anisotropy in the upper mantle implies that the temperature distribution, rheology, and, hence, the upper-mantle dynamics depend on its deformation history. Olivine orientations frozen in the continental lithosphere may modify plume–lithosphere interactions. Enhanced thermal diffusivity along lithospheric-scale wrench zones—that is, parallel to the olivine [100] preferred orientation—may lead to anisotropic heating of the lithosphere, favouring the reactivation of these structures during continental break-up[23,24]. Such control of the pre-existing lithospheric structure on the propagation of a thermal anomaly may be inferred, for instance, from tomographic images of the east African rift[25]. In oceanic domains, deformation by resistive drag of the sublithospheric mantle leads to vertical orientation of olivine [010] axes in the lithosphere and asthenosphere[6], which will result in reduced vertical diffusivity and hence in less effective heat transfer from the convective mantle. Reduced diffusivity across oceanic plates may also lead to slower heating of subducting slabs. Interactions between flow and anisotropy are nevertheless complex and a full investigation of the effect of an anisotropic thermal diffusivity on mantle convection may only be accomplished by integrating this strain-induced anisotropy in geodynamical models. □

Methods

Lattice-preferred orientation measurements

Crystallographic orientations were determined on a scanning electron microscope (JEOL JSM 5600) by indexation of electron back-scattered diffraction (EBSD) patterns generated by interaction of a vertical incident electron beam with a carefully polished thin section tilted at 70°. The pattern is projected onto a phosphor screen and recorded by a low-light, high-resolution charge-coupled device (CCD) camera. The image is then digitally processed and indexed in terms of crystal orientation using the CHANNEL+ software from HKL Technology.

Thermal diffusivity measurements

For each sample, thermal diffusivity of one (PNG) to six (BALD1) oriented cores of 43 mm length and 27 mm diameter was determined using a finite pulse method[26]. The energy pulse (3 s) is generated at the core axis by a Nichrome heater. It propagates radially and is recorded by a Chromel/Alumel thermocouple placed at a distance d (6–7 mm) from the core axis. The temperature response is measured every 0.1 s during 40 s. From the measured temperature curve, the half-time value t_H, that is, the time span to reach half of the maximum temperature rise, is evaluated. Thermal diffusivity K is then calculated as $K = d^2/(10.77t_H - 16.55)$. A thick-walled pyrophyllite container prevents thermal convection around the sample and a drastic step in thermal properties at sample boundaries. The entire assemblage is fitted in a three-zone furnace. At each temperature, three successive measurements are performed in order to check data reproducibility. Moreover, for samples BALM4 and BALD1, two series of measurements (two cores) were performed in each direction.

Received 17 October 2000; accepted 25 April 2001.

1. Kobayashi, Y. Anisotropy of thermal diffusivity in olivine, pyroxene, and dunite. *J. Phys. Earth* **22,** 359–373 (1974).
2. Chai, M., Brown, J. M. & Slutsky, L. J. Thermal diffusivity of mantle minerals. *Phys. Chem. Minerals* **23,** 470–475 (1996).
3. Silver, P. G. Seismic anisotropy beneath the continents: Probing the depths of geology. *Annu. Rev. Earth Planet. Sci.* **24,** 385–432 (1996).
4. Nicolas, A. & Christensen, N. I. in *Composition, Structure and Dynamics of the Lithosphere–Asthenosphere System* (eds Fuchs, K. & Froidevaux, C.) 111–123 (American Geophysical Union, Washington DC, 1987).
5. Schärmeli, G. H. in *High-Pressure Researches in Geosciences* (ed. Schreyer, W.) 349–373 (E. Schweizerbart'sche, Stuttgart, 1982).
6. Tommasi, A. Forward modeling of the development of seismic anisotropy in the upper mantle. *Earth Planet. Sci. Lett.* **160,** 1–13 (1998).
7. Chastel, Y. B., Dawson, P. R., Wenk, H.-R. & Bennet, K. Anisotropic convection with implications for the upper mantle. *J. Geophys. Res.* **98,** 17757–17771 (1993).
8. Blackman, D. K. *et al.* Teleseismic imaging of subaxial flow at mid-ocean ridges: travel-time effects of anisotropic mineral texture in the mantle. *Geophys. J. Int.* **127,** 415–426 (1996).
9. Tommasi, A., Mainprice, D., Canova, G. & Chastel, Y. Viscoplastic self-consistent and equilibrium-based modeling of olivine lattice preferred orientations. Implications for upper mantle seismic anisotropy. *J. Geophys. Res.* **105,** 7893–7908 (2000).
10. Mainprice, D. & Humbert, M. Methods of calculating petrophysical properties from lattice preferred orientation data. *Surv. Geophys.* **15,** 575–592 (1994).
11. Kanamori, H., Fuji, N. & Mizutani, H. Thermal diffusivity measurement of rock-forming minerals from 300 °K to 1100 °K. *J. Geophys. Res.* **73,** 595–605 (1968).
12. Kasahara, J., Suzuki, I., Kumazawa, M., Kobayashi, Y. & Iida, K. Anisotropism of P wave in dunite. *Zisin (J. Seismol. Soc. Jpn)* **21,** 222–228 (1968).
13. Horai, K. & Susaki, J. The effect of pressure on the thermal conductivity of silicate rocks up to 12 kbar. *Phys. Earth Planet. Inter.* **55,** 292–305 (1989).
14. Seipold, U., Mueller, H. J. & Tuisku, P. Principle differences in the pressure dependence of thermal and elastic properties of crystalline rocks. *Phys. Chem. Earth* **23,** 357–360 (1998).
15. Katsura, T. Thermal diffusivity of olivine under upper mantle conditions. *Geophys. J. Int.* **122,** 63–69 (1995).
16. Shankland, T. J., Nitsan, U. & Duba, A. G. Optical absorption and radiative heat transport in olivine at high temperature. *J. Geophys. Res.* **84,** 1603–1610 (1979).

© 2001 Macmillan Magazines Ltd

17. Schatz, J. F. & Simmons, G. Thermal conductivity of Earth materials at high temperatures. *J. Geophys. Res.* **77,** 6966–6983 (1972).
18. Holt, J. B. Thermal diffusivity of olivine. *Earth Planet. Sci. Lett.* **27,** 404–408 (1975).
19. Seipold, U. Temperature dependence of thermal transport properties of crystalline rocks–general law. *Tectonophysics* **291,** 161–171 (1998).
20. Hofmeister, A. M. Mantle values of thermal conductivity and the geotherm from phonon lifetimes. *Science* **283,** 1699–1706 (1999).
21. Klemens, P. G. Thermal conductivity and lattice vibrational modes. *Solid State Phys.* **7,** 1–98 (1958).
22. Bouhifd, M. A., Andrault, D., Fiquet, G. & Richet, P. Thermal expansion of forsterite up to the melting point. *Geophys. Res. Lett.* **23,** 134–1136 (1996).
23. Vauchez, A., Barruol, G. & Tommasi, A. Why do continents break up parallel to ancient orogenic belts? *Terra Nova* **9,** 62–66 (1997).
24. Tommasi, A. & Vauchez, A. Continental rifting parallel to ancient collisional belts: An effect of the mechanical anisotropy of the lithospheric mantle. *Earth Planet. Sci. Lett.* **185,** 199–210 (2001).
25. Achauer, U. & KRISP working group. New ideas on the Kenya rift based on the inversion of the combined dataset of the 1985 and 1989/90 seismic tomography experiments. *Tectonophysics* **236,** 305–329 (1988).
26. Seipold, U. Simultaneous measurements of thermal diffusivity and thermal conductivity under high pressure using thermal pulse of finite length. *High Temp.-High Press.* **20,** 609–613 (1988).
27. Abramson, E. H., Brown, M., Slutsky, L. J. & Zaug, J. The elastic constants of San Carlos olivine up to 17 GPa. *J. Geophys. Res.* **102,** 12252–12263 (1997).
28. Duffy, T. S. & Vaughan, M. T. Elasticity of enstatite and its relationships to crystal structure. *J. Geophys. Res.* **93,** 383–391 (1988).

Acknowledgements

We thank C. Karger for participation in the measurements, and M. Ferrero and F. Boudier for providing the samples from Baldissero and Papua New Guinea, respectively. The Laboratoire de Tectonophysique's EBSD system was funded by the CNRS/INSU, Université of Montpellier II, and NSF project "Anatomy of an archean craton". This work was supported by the CNRS/INSU programme "Action Thématique Innovante".

Correspondence should be addressed to A.T.
(e-mail: deia@dstu.univ-montp2.fr).

Proteorhodopsin phototrophy in the ocean

Oded Béjà*†, Elena N. Spudich†‡, John L. Spudich‡, Marion Leclerc* & Edward F. DeLong*

* *Monterey Bay Aquarium Research Institute, Moss Landing, California 95039, USA*
‡ *Department of Microbiology and Molecular Genetics, The University of Texas Medical School, Houston, Texas 77030, USA*
† *These authors contributed equally to this work*

Proteorhodopsin[1], a retinal-containing integral membrane protein that functions as a light-driven proton pump, was discovered in the genome of an uncultivated marine bacterium; however, the prevalence, expression and genetic variability of this protein in native marine microbial populations remain unknown. Here we report that photoactive proteorhodopsin is present in oceanic surface waters. We also provide evidence of an extensive family of globally distributed proteorhodopsin variants. The protein pigments comprising this rhodopsin family seem to be spectrally tuned to different habitats—absorbing light at different wavelengths in accordance with light available in the environment. Together, our data suggest that proteorhodopsin-based phototrophy is a globally significant oceanic microbial process.

Bacteriorhodopsin, a retinal-containing membrane protein that functions as a light-driven proton pump, was discovered nearly three decades ago in the halophile *Halobacterium salinarum*[2]. The 'purple membranes' of this archaeon contain a high concentration of bacteriorhodopsin molecules, packed in an ordered two-dimensional crystalline array[3]. On absorption of light, bacteriorhodopsin undergoes a series of conformational shifts (a photocycle), causing a proton to be transported across the membrane. The resulting electrochemical membrane potential drives ATP synthesis, through an H^+-ATPase[4]. Similar rhodopsin-mediated, light-driven proton pumping, formerly thought to exist only in halophilic archaea, has been discovered in an uncultivated marine bacterium[1] of the 'SAR86' phylogenetic group[5]. This finding suggested that a previously unrecognized phototrophic pathway may occur in the ocean's photic zone; however, all earlier data are based solely on recombinant DNA and *in vitro* biochemical analyses, and this phenomenon has not yet been observed in the sea.

To test whether rhodopsin-like molecules form photoactive proteins in native marine bacteria, we analysed membrane preparations of bacterioplankton collected directly from Monterey Bay surface waters. Using laser flash-photolysis techniques[6], we searched for spectroscopic evidence of proteorhodopsin-like photochemical activity in these bacterioplankton membrane preparations. We observed transient flash-induced absorption changes attributable to a photochemical reaction cycle of 15 ms (Fig. 1), which is characteristic of retinylidene ion pumps[7] and similar to the photocycle seen with *Escherichia coli* membranes expressing recombinant proteorhodopsin[1].

Furthermore, 5 ms after the flash the absorption difference spectrum of the environmental sample exhibited the same shape and same isosbestic point for the light and dark state as proteorhodopsin expressed in *E. coli* membranes. This shows that photochemical reactions in native bacterioplankton cell membranes generate a red-shifted, long-lived intermediate spectrally similar to that of recombinantly expressed proteorhodopsin in *E. coli*. The flash-photolysis data provide direct physical evidence for the existence of proteorhodopsin-like transporters and endogenous retinal molecules in the microbial fraction of these coastal surface waters.

The photocycling pigment was bleached with hydroxylamine and, after hydroxylamine removal, photoactivity was restored by adding all-*trans* retinal (Fig. 2), showing that the photosignals do indeed derive from retinylidene pigmentation. We conclude that

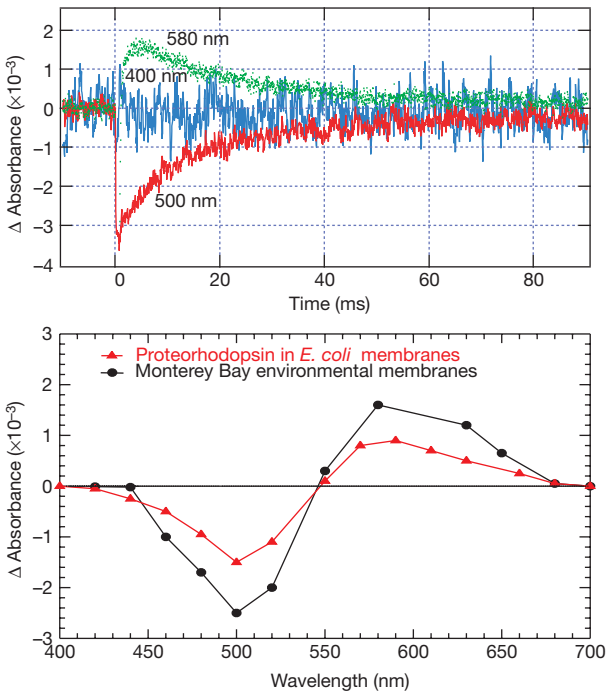


Figure 1 Laser flash-induced absorbance changes in suspensions of membranes prepared from the prokaryotic fraction of Monterey Bay surface waters. Top, membrane absorption was monitored at the indicated wavelengths and the flash was at time 0 at 532 nm. Bottom, absorption difference spectrum at 5 ms after the flash for the environmental sample (black) and for *E. coli*-expressed proteorhodopsin (red).

© 2001 Macmillan Magazines Ltd

proteorhodopsin molecules are present in the membranes of native marine bacterioplankton.

The flash-photolysis data permit estimation of the cellular concentration of proteorhodopsin. Assuming (1) the flash yield of the environmental pigment is similar to that of proteorhodopsin expressed in *E. coli* (that is, 10.5% absorption change at 500 nm per absorption unit of pigment at 527 nm), we calculate 0.033 absorption units of proteorhodopsin at 527 nm (0.35 μg of proteorhodopsin protein per litre of sea water). Flash spectroscopy was necessary to detect the proteorhodopsin pigments because there is much greater absorption in the visible range by other pigments in the sample, which show peaks of 1.6, 1.3, 0.52 and 0.73 absorption units, at 437, 468, 647 and 673 nm, respectively.

Assuming also (2) a molar absorption coefficient of 50,000 M^{-1} cm^{-1} at the absorption maximum, (3) fluorescence *in situ* hybridization counts of a total of 5.6×10^{10} SAR86 cells in the concentrated sample (~10% of the total bacteria; see Methods), (4) 50% recovery of membranes from the cells, and (5) that the uncultivated SAR86 group is the principal bacterioplankton group using these pigments, we calculate that there are 2.4×10^4 proteorhodopsin molecules per SAR86 cell.

This value is on the order of the concentration of bacteriorhodopsin in a *H. salinarum* cell, in which substantial portions of the membrane surface area of the cell consist of bacteriorhodopsin in a tightly packed crystalline array[8]. For comparison, 2.4×10^4 bacteriorhodopsin molecules densely packed in the purple-membrane lattice would cover a 0.6-μm diameter, flat circular area of membrane—a significant portion of the surface of a single cell[3]. This number of molecules is sufficient to produce substantial amounts of ATP under illumination[9]. Therefore, the high density of proteorhodopsin in the SAR86 membrane indicated by our calculations strongly suggests that this protein has a significant role in the physiology of these bacteria *in situ*.

To explore the possible existence of other proteorhodopsins, we screened the same mixed-population bacterial artificial chromosome (BAC) library[10] in which proteorhodopsin was initially discovered, with non-degenerate polymerase chain reaction (PCR) primers[1]. Several additional proteorhodopsin-containing BAC clones were found in the library. These proteorhodopsins were similar, but did differ in their amino-acid sequences when compared with the original proteorhodopsin (for example, see clones 31A8, 64A5 and 40E8; Figs 3 and 4). Using the same non-degenerate primers, we could also amplify by PCR proteorhodopsin genes from bacterioplankton DNA extracts, including those from Monterey Bay (MB clones; Fig. 3), the Southern ocean (Palmer station; PAL clones) and waters of the central North Pacific ocean (Hawaii Ocean Time series station[11]; HOT clones).

We detected 15 different variants of proteorhodopsin in the PCR-generated Monterey Bay proteorhodopsin gene library, falling into three clusters (Fig. 3) that share at least 97% identity over 248 amino acids (Fig. 4) and 93% identity at the DNA level. Two proteorhodopsin genes from Monterey, 40E8 and 64A5, were expressed in *E. coli* and produced absorption spectra very similar to the original proteorhodopsin[1] (clone 31A8) isolated from the same waters (data not shown).

Remarkably, all the proteorhodopsin genes that were amplified by PCR from Antarctic marine bacterioplankton were different from those of Monterey Bay (Fig. 3), sharing a maximum of 78% identity over 248 amino acids with the Monterey clade (Fig. 4). The changes in amino-acid sequences were not restricted to the hydrophilic

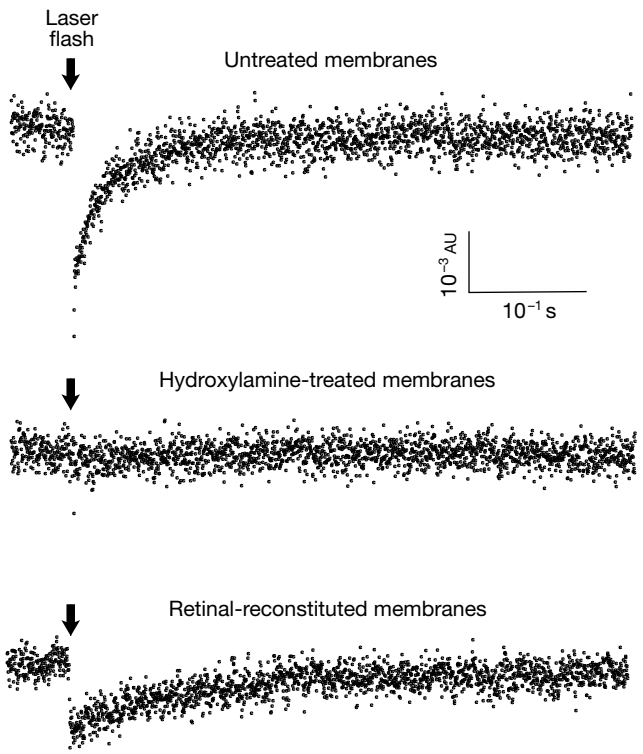


Figure 2 Laser flash-induced transients at 500 nm of a Monterey Bay bacterioplankton membrane preparation. Top, before addition of hydroxylamine; middle, after 0.2 M hydroxylamine treatment at pH 7.0, 18 °C, with 500-nm illumination for 30 min; bottom, after centrifuging twice with resuspension in 100 mM phosphate buffer, pH 7.0, followed by addition of 5 μM all-*trans* retinal and incubation for 1 h.

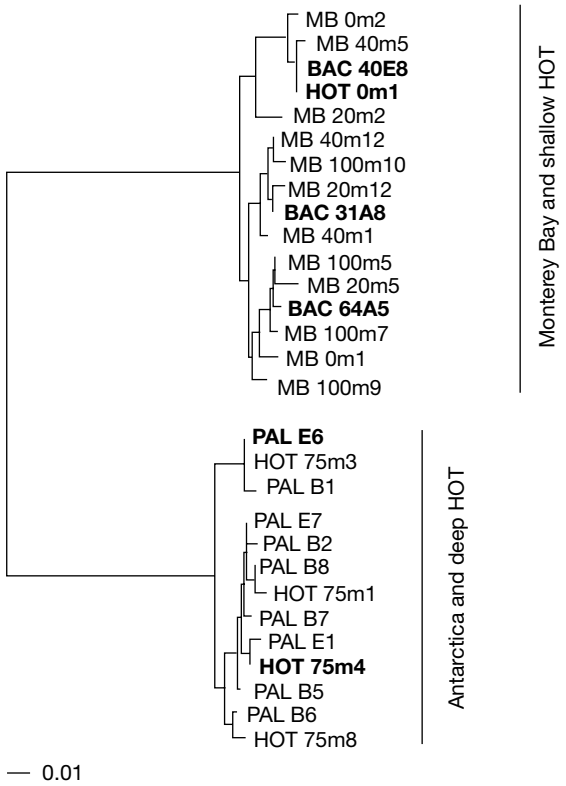


Figure 3 Phylogenetic analysis of the inferred amino-acid sequence of cloned proteorhodopsin genes. Distance analysis of 220 positions was used to calculate the tree by neighbour-joining using the PaupSearch program of the Wisconsin Package version 10.0 (Genetics Computer Group; Madison, Wisconsin). *H. salinarum* bacteriorhodopsin was used as an outgroup, and is not shown. Scale bar represents number of substitutions per site. Bold names indicate the proteorhodopsins that were spectrally characterized in this study.

© 2001 Macmillan Magazines Ltd

loops but were spread over the entire protein, including changes near the retinal-binding domain (Fig. 4). Also, there was an insertion of one amino acid in the Antarctic proteorhodopsins, relative to those of the Monterey clade (Fig. 4).

We expressed the palE6 proteorhodopsin gene from Antarctica in *E. coli* cells. Addition of retinal to membranes of cells containing Antarctic proteorhodopsin apoprotein produced an absorbance peak at 490 nm (Fig. 5)—a blue shift of 37 nm from the 527-nm peak observed for the Monterey Bay proteorhodopsins. The Antarctic proteorhodopsin spectrum exhibited vibrational fine structure, as observed previously in retinylidene pigments for archaeal sensory rhodopsin II, which has a very similar blue-shifted (487-nm) absorption maximum[12]. The nearly identical shapes of the structured spectra suggest similar mechanisms of wavelength regulation in the bacterial and archaeal rhodopsins. Furthermore, palE6 proteorhodopsin functions similarly to its Monterey Bay homologues[1], exhibiting light-mediated transport of protons in right-side-out *E. coli* vesicles (E. N. Spudich *et al.*, manuscript in preparation).

We screened a fosmid library from Antarctica (E. F. DeLong, unpublished data) with proteorhodopsin primers, and sequenced one clone containing the gene. Preliminary sequence analyses based on flanking gene order and identity indicate that, despite differences in amino-acid sequence and absorption spectra, the Antarctic proteorhodopsin is derived from a bacterium highly related to the proteorhodopsin-containing SAR86 bacteria of Monterey Bay[10] (O. Béjà, unpublished data).

Thirty-two proteorhodopsin genes from the HOT station were cloned from surface (5-m) and 75-m samples. Most proteorhodopsin clones (80%) from the HOT surface waters belonged to the Monterey clade and were identical (on the amino-acid level) to the proteorhodopsin gene from BAC 40E8. In contrast, most of the clones from the 75-m sample (90%) fell within the Antarctic clade (Fig. 3). Two proteorhodopsin genes from the HOT station, HOT 0m1 and HOT 75m4, were expressed in *E. coli* and their absorption spectra were examined. As expected from their primary sequences, the HOT proteorhodopsin clustering with the Antarctic clade gave a blue (490-nm) absorption maximum, whereas the proteorhodopsin clustering with the Monterey clade gave a green (527-nm) absorption maximum (Fig. 5).

In the oligotrophic waters of the central North Pacific gyre, most of the light energy is in the blue range, with maximal intensity near 475 nm (ref. 13). This energy peak is maintained over depth, whereas the total energy decreases. At the surface, the energy peak is very broad with a half bandwidth between 400 and 650 nm. In deeper water below 50 m, the peak narrows and the half bandwidth is restricted to between 450 and 500 nm (Fig. 5). Considering the different wavelengths absorbed by members of the two different proteorhodopsin clades (527 nm versus 490 nm; Fig. 5), it seems that the blue-shifted proteorhodopsin variants are better adapted to the light available in their environment. Notably, rod and cone visual pigment rhodopsins from closely related fish species in Lake Baikal[14] also exhibit a blue shift in deeper dwelling species. In addition, previous studies of cultivated *Prochlorococcus*[15,16] and *Synechococcus*[17], as well as natural cyanobacterial communities[18], also indicate the presence of surface- and deep-water groups adapted to different depths.

Our data now show the presence of co-existing surface- and deep-water clades of proteorhodopsin-based phototrophs, whose energy-generating pigments are spectrally tuned to either shallow or deeper water light fields. Presumably, this provides subsets of proteorhodopsin genetic variants in mixed populations with a selective advantage at different points along the depth-dependent light

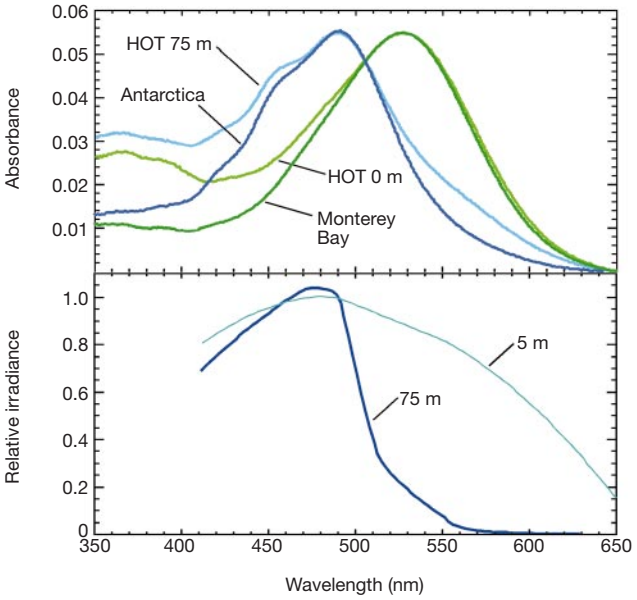


Figure 5 Absorption spectra of retinal-reconstituted proteorhodopsins in *E. coli* membranes. All-*trans* retinal (2.5 μM) was added to membrane suspensions in 100 mM phosphate buffer, pH 7.0, and absorption spectra were recorded. Top, four spectra for palE6 (Antarctica), HOT 75m4, HOT 0m1, and BAC 31A8 (Monterey Bay) at 1 h after retinal addition. Bottom, downwelling irradiance from HOT station measured at six wavelengths (412, 443, 490, 510, 555 and 665 nm) and at two depths, for the same depths and date that the HOT samples were collected (0 and 75 m). Irradiance is plotted relative to irradiance at 490 nm.

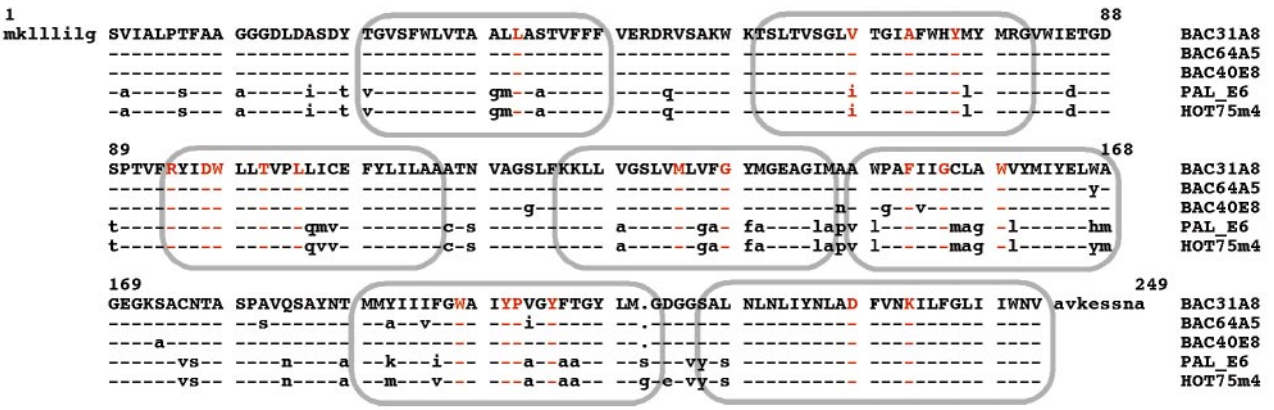


Figure 4 Multiple alignment of proteorhodopsin amino-acid sequences. The secondary structure shown (boxes for transmembrane helices) is derived from hydropathy plots. Residues predicted to form the retinal-binding pocket are marked in red.

© 2001 Macmillan Magazines Ltd

gradient. Of course, genetic variation other than proteorhodopsin tuning is likely to influence the fitness of population variants along the depth gradient. Other genetic loci may also co-vary with tuned absorption maxima, indicating a depth-optimized adaptation for these other traits as well. Genomic approaches should provide a means to test such hypotheses.

Proteorhodopsin-mediated phototrophic energy generation may have a significant impact on carbon and energy flux in the ocean. Although it is not known whether proteorhodopsin-bearing bacteria fix CO_2, their ability to generate energy from light should reduce overall respiratory energy requirements, as has been suggested for aerobic, bacteriochlorophyll-containing marine bacteria[19]. Bioenergetic calculations based on our data suggest that proteorhodopsin-based phototrophy could provide a large fraction of energy requirements for cell maintenance and reproduction. The widespread distribution and high abundance of SAR86 bacteria[5,10,20–22], combined with the biophysical and genetic data that we report, indicate that proteorhodopsin phototrophy is a biologically significant, globally distributed marine microbial process. □

Methods

Cell collection and membrane preparation

Antarctic coastal waters were collected off Palmer Station (64.4° S, 64.0° W), Anvers Island, Antarctic Peninsula, in August 1996 (ref. 23). Hawaiian HOT station (22.4° N, 158.0° W) waters were collected on 17 March 1998. Cell membranes were prepared from picoplankton cells concentrated from surface seawater (700 l) collected in early winter 2000, roughly 26 miles offshore from Moss Landing, California (36.7° N, 122.4° W). Water samples were prefiltered through a GF/A glass-fibre filter (approximate particle size $< 0.6\ \mu m$) to remove the larger eukaryotic phytoplankton cells. We concentrated the filtrate by tangential flow filtration[10] and prepared the membranes[24] as described. Aliquots of the cell preparation were fixed in 3.7% formalin for subsequent fluorescence *in situ* hybridization and taxon-specific cell quantification.

Fluorescence *in situ* hybridization

BAC clones 31A08 and 27G05 (ref. 10) were used as templates to generate SAR86-specific polyribonucleotide probes targeted to large-subunit ribosomal RNA. We prepared fluorescently labelled polyribonucleotide probes that targeted a variable region of large subunit rRNA by *in vitro* transcription (M. Leclerc *et al*., manuscript in preparation), and carried out subsequent whole-cell hybridization assays as described[25]. The percentage of SAR86-type cells relative to total 4′,6-diamidino-2-phenylindole dihydrochloride (DAPI)-stained cells was calculated as described[25].

Spectroscopy

One-half of the cell-membrane suspension (corresponding to 350 l of seawater) was suspended in 2 ml of 100 mM sodium phosphate buffer, pH 7.0. Absorption spectra were measured on an Aminco DW2000 UV-visible absorption spectrophotometer (SLM; Urbana, IL). Flash-induced absorption transients in the millisecond to seconds time domain were acquired on a digital oscilloscope (Nicolet, Integra20) after a Nd-YAG laser flash (532 nm, 6-ns pulse duration, 40 mJ; Continuum, Surelight I) in a laboratory-constructed flash-photolysis system as described[6]. Each transient was obtained by averaging 64 acquisition sweeps collected at 30-s intervals at various wavelengths between 360 and 680 nm. We performed kinetic analyses with the program IGOR Pro, v3.1 (WaveMetrics, Lake Oswego, OR).

Received 3 January; accepted 26 March 2001.

1. Béjà, O. *et al.* Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289,** 1902–1906 (2000).
2. Oesterhelt, D. & Stoeckenius, W. Rhodopsin-like protein from the purple membrane of *Halobacterium halobium*. *Nature* **233,** 149–152 (1971).
3. Henderson, R. & Unwin, P. N. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257,** 28–32 (1975).
4. Danon, A. & Stoeckenius, W. Photophosphorylation in *Halobacterium halobium*. *Proc. Natl Acad. Sci. USA* **71,** 1234–1238 (1974).
5. Mullins, T. D., Britcshgi, T. B., Krest, R. L. & Giovannoni, S. J. Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnol. Oceanogr.* **40,** 148–158 (1995).
6. Sasaki, J. & Spudich, J. L. The transducer protein HtrII modulates the lifetimes of sensory rhodopsin II photointermediates. *Biophys. J.* **75,** 2435–2440 (1998).
7. Spudich, J. L., Yang, C. S., Jung, K. H. & Spudich, E. N. Retinylidene proteins: Structures and functions from archaea to humans. *Annu. Rev. Cell Dev. Biol.* **16,** 365–392 (2000).
8. Neugebauer, D. C., Zingsheim, H. P. & Oesterhelt, D. Biogenesis of purple membrane in halobacteria. *Methods Enzymol.* **97,** 218–226 (1983).
9. Michel, H. & Oesterhelt, D. Electrochemical proton gradient across the cell membrane of *Halobacterium halobium*: effect of *N,N′*-dicyclohexylcarbodiimide, relation to intracellular adenosine triphosphate, adenosine diphosphate, and phosphate concentration, and influence of the potassium gradient. *Biochemistry* **19,** 4607–4614 (1980).
10. Béjà, O. *et al.* Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2,** 516–529 (2000).
11. Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep-Sea Res.* **43,** 129–156 (1996).
12. Takahashi, T. *et al.* Color regulation in the archaebacterial phototaxis receptor phoborhodopsin (sensory rhodopsin II). *Biochemistry* **29,** 8467–8474 (1990).
13. Kirk, J. T. O. in *Light and Photosynthesis in Aquatic Ecosystems* 104–134 (Cambridge Univ. Press, Cambridge, 1983).
14. Bowmaker, J. K. The ecology of visual pigments. *Novartis Found. Symp.* **224,** 21–31 (1999).
15. Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). *J. Mol. Evol.* **46,** 188–201 (1998).
16. Moore, L. R., Rocap, G. & Chisholm, S. W. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393,** 464–467 (1998).
17. Palenik, B. Chromatic adaptation in marine *Synechococcus* strains. *Appl. Environ. Microbiol.* **67,** 991–994 (2001).
18. Ferris, M. J. & Palenik, B. Niche adaptation in ocean cyanobacteria. *Nature* **396,** 226–228 (1998).
19. Kolber, Z. S., Van Dover, C. L., Niderman, R. A. & Falkowski, P. G. Bacterial photosynthesis in surface waters of the open ocean. *Nature* **407,** 177–179 (2000).
20. Eilers, H., Pernthaler, J., Glockner, F. O. & Amann, R. Culturability and *in situ* abundance of pelagic bacteria from the North Sea. *Appl. Environ. Microbiol.* **66,** 3044–3051 (2000).
21. Gonzalez, J. M. *et al.* Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* **66,** 4237–4246 (2000).
22. Rappé, M. S., Vergin, K. & Giovannoni, S. J. Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. *FEMS Microbiol. Ecol.* **33,** 219–232 (2000).
23. DeLong, E. F. *et al.* Dibiphytanyl ether lipids in nonthermophilic crenarchaeotes. *Appl. Environ. Microbiol.* **64,** 1133–1138 (1998).
24. Shimono, K., Iwamoto, M., Sumi, M. & Kamo, N. Functional expression of pharaonis phoborhodopsin in *Escherichia coli*. *FEBS Lett.* **420,** 54–56 (1997).
25. DeLong, E. F., Taylor, L. T., Marsh, T. L. & Preston, C. M. Visualization and enumeration of marine planktonic archaea and bacteria by using polyribonucleotide probes and fluorescent *in situ* hybridization. *Appl. Environ. Microbiol.* **65,** 5554–5563 (1999).

Acknowledgements

We thank J. Zehr and D. Karl for the HOT samples and the captain and crew of the RV *Point Lobos* for expert assistance at sea. D. Karl and R. Letellier provided spectral irradiance data. This research was supported by the National Science Foundation (E.F.D.), the NIH (J.L.S.), and the David and Lucile Packard Foundation (E.F.D.).

Correspondence and requests for materials should be addressed to E.F.D. (e-mail: delong@mbari.org). The sequences have been deposited with GenBank under accession numbers AF349976–AF350003.

Large-scale forest girdling shows that current photosynthesis drives soil respiration

Peter Högberg*, Anders Nordgren*, Nina Buchmann†, Andrew F. S. Taylor‡, Alf Ekblad*§, Mona N. Högberg*, Gert Nyberg*, Mikaell Ottosson-Löfvenius* & David J. Read‖

** Section of Soil Science, Department of Forest Ecology, SLU, SE-901 83 Umeå, Sweden*
† Max-Planck Institute for Biogeochemistry, PO Box 100164, 07701 Jena, Germany
‡ Department of Forest Mycology and Pathology, SLU, PO Box 7026, SE-750 07 Uppsala, Sweden
‖ Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

The respiratory activities of plant roots, of their mycorrhizal fungi and of the free-living microbial heterotrophs (decomposers) in soils are significant components of the global carbon balance, but their relative contributions remain uncertain[1,2]. To separate mycorrhizal root respiration from heterotrophic respiration in a

§ Present address: Department of Natural Sciences, Örebro University, SE-701 82 Örebro, Sweden.

© 2001 Macmillan Magazines Ltd

boreal pine forest, we conducted a large-scale tree-girdling experiment, comprising 9 plots each containing about 120 trees. Tree-girdling involves stripping the stem bark to the depth of the current xylem at breast height terminating the supply of current photosynthates to roots and their mycorrhizal fungi without physically disturbing the delicate root–microbe–soil system. Here we report that girdling reduced soil respiration within 1–2 months by about 54% relative to respiration on ungirdled control plots, and that decreases of up to 37% were detected within 5 days. These values clearly show that the flux of current assimilates to roots is a key driver of soil respiration; they are conservative estimates of root respiration, however, because girdling increased the use of starch reserves in the roots. Our results indicate that models of soil respiration should incorporate measures of photosynthesis and of seasonal patterns of photosynthate allocation to roots.

Flux measurements suggest that boreal forests can be either sources or sinks for atmospheric carbon dioxide (CO_2; refs 3, 4), and that there is considerable interannual variability in this respect[5,6]. The forest carbon (C) balance is the net result of CO_2 fixation by photosynthesis occurring above ground and the release of C as CO_2, notably from the below-ground compartment through the respiratory activities of plant roots, their symbiotic mycorrhizal fungi and the free-living microbial and faunal populations of the soil. Of the two compartments, the below-ground system is the most difficult to evaluate[2].

Almost all approaches to its study have used methods that physically disturb the intimately linked processes in which C is allocated to fine roots, their mycorrhizal symbionts, and through these to the wider soil community. Estimates of the contribution of root respiration to total soil respiration vary from 10 to 90% (ref. 2). Although some of this variability reflects differences among types of ecosystems, a considerable proportion is probably caused by problems of experimental design or methodology. As a result, our knowledge of the absolute and relative contributions of different components of the root–heterotroph–soil continuum to the forest C balance is limited[2].

The soil remains a 'black box' despite the need to obtain accurate models of this vital ecosystem compartment. This gap in our understanding becomes particularly serious in the context of boreal forests—one of the two most extensive terrestrial biomes on earth, which accounts, together with adjacent ecosystems of high latitudes, for the largest global capital of soil organic C (ref. 7). The identification of factors that control C source–sink relations is a prerequisite for understanding the boreal ecosystem C balance and the potential effects of raised temperatures, atmospheric CO_2 concentrations and deposition of nitrogen.

To determine the main drivers of soil respiration and to evaluate more precisely the relative contributions to CO_2 efflux made by the different biota, we conducted a large-scale stem-girdling experiment in a boreal Scots pine (*Pinus sylvestris* L.) forest in northern Sweden. Girdling has the instantaneous effect of terminating the flux of photosynthates from the tree canopy through the phloem to the roots, while enabling water transport in the reverse direction through the xylem. Thus, girdling has no immediate impact on the principal physical soil parameters, moisture and temperature, nor does it physically displace roots and soil organisms or sever roots and fungal hyphae.

Retention of the integrity of the root–fungus pathway is especially important in ecosystems of this type, in which the surfaces of virtually all roots are ensheathed by a thick mantle of fungal hyphae to form ectomycorrhizas[8]. The extensive but inherently delicate mycelial system extending from these symbiotic structures into the soil not only scavenges for mineral nutrients required by the trees, but also provides the conduits through which the C passes to enable the fungal sexual structures—the sporocarps—to develop.

To indicate the potential of influx of C from ungirdled trees into plots with girdled trees, we examined the distribution of ectomycorrhizal sporocarps. To avoid edge effects, which are inevitably associated with small-scale girdling experiments, our plots covered large (900 m^2) areas. Soil respiration was measured within a metre of the centre of each plot. All trees in six such large plots, with about 120 trees in each plot, were girdled either in early June (early girdled) or in August (late girdled). This duplication was designed

Table 1 Sporocarps of ectomycorrhizal fungi and starch in fine roots

(a) Ectomycorrhizal fungi*

Treatment	No. of species	No. of sporocarps	Biomass (g dry wt)
Control	11 ± 3	252 ± 116	84.4 ± 24.4
Early girdling	1 ± 1 (72)	4 ± 4 (72)	0.5 ± 0.5 (72)
Late girdling	10 ± 1 (2–3)	108 ± 48 (2–3)	40.9 ± 8.5 (2–3)

(b) Concentrations of starch in fine roots†

	Starch (% of dry wt)		
Treatment	July 6	August 17	October 16
Control	10.3 ± 1.3	4.6 ± 0.5	5.1 ± 0.6
Early girdling	3.5 ± 1.4 (28)	1.0 ± 0.3 (68)	0.6 ± 0.1 (130)
Late girdling	ND	5.0 ± 0.5	1.5 ± 0.4 (59)

Data are means ± s.e.m. (n = 3) per plot. Figures in parentheses give the number of days since the girdling treatment.

* The number of species, the number of sporocarps and the biomass of ectomycorrhizal fungi on the central 100 m^2 of the experimental plots in the different treatments in the tree girdling experiment at Åheden.

† Concentrations of starch in fine (< 2 mm in diameter) roots of *P. sylvestris*.

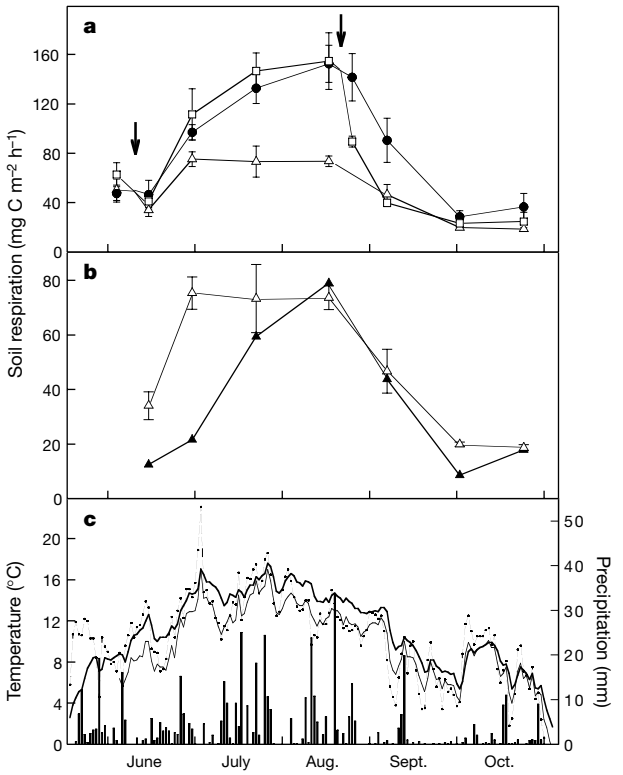


Figure 1 Soil respiration in the different tree-girdling treatments in Scots pine forest at Åheden in relation to meteorological data. **a**, Respiratory soil CO_2 efflux from ungirdled control (filled circles), early girdled (open triangles) and late girdled (open squares) plots. **b**, Calculated root respiration (filled triangles, respiration on control plots minus respiration on early girdled plots) and heterotrophic respiration (open triangles, respiration on early girdled plots). **c**, Open-field meteorological station data for precipitation (bars), daily averages of air temperature and soil temperature at 5-cm soil depth in the mineral soil (broken and thick solid lines, respectively), and average soil temperatures at 4-cm depth below the mor layer but above the mineral soil (thin solid line), in the studied forest. Arrows indicate times of early and late girdling.

© 2001 Macmillan Magazines Ltd

to allow the detection of any phenological effects on respiratory activity. Three plots of equivalent size were established in adjacent areas but left ungirdled as controls. The respiratory activity remaining on girdled plots was ascribed to heterotrophic activity, whereas the difference in respiration between control and girdled plots was ascribed to respiration by mycorrhizal roots. Concentrations of starch were measured to determine whether girdling resulted in increased use of stored carbohydrates in roots.

Two months after the initiation of the early girdling, comparisons of the numbers and biomass of ectomycorrhizal fungal fruit bodies in the central 100 m^2 of the early girdled and control plots showed that these structures had been virtually eliminated by the girdling procedure (Table 1a). An increasing abundance of sporocarps closer to the edge of treated plots (data not shown) suggested an influence of untreated trees outside treated plots. These observations supported the use of 900-m^2 plots, and especially the use of the centre of the plots for measurements of treatment effects on soil respiration. The observations also support the view, based mainly on experiments carried out in laboratory microcosms (for example, see ref. 9), that the ectomycorrhizal mycelial system is strongly dependent on current assimilates. In contrast, there was no reduction in the numbers or biomass of saprotrophic fungi in the girdled relative to the control plots (data not shown).

Soil respiration on the control plots followed a seasonal pattern with rates of 20–60 mg CO_2–C m^{-2} h^{-1} early and late in the growing season, and with a maximum of 150 mg CO_2–C m^{-2} h^{-1} in mid-August (Fig. 1a). In the early girdled plots, soil respiration showed within 5 days a decrease of 27% relative to that measured in control plots. This relative difference between treatments was retained for a further 2 weeks, although soil respiration increased in both treatments. By this time, at the end of June, respiration in the early girdling treatment had reached its maximum. Respiration on control plots continued to increase for a further 7 weeks until mid-August, indicating that the increase was due to root rather than heterotrophic activity. This resulted in a 52% lower respiration on early girdled plots compared with on control plots. The late girdling produced even faster effects; soil respiration declined by 37% relative to control plots within 5 days of treatment (Fig. 1a). This precipitous drop in activity continued such that by 2 weeks the CO_2 efflux was 56% lower than in control plots. As a result, the respiratory activity of the late girdled plots rapidly reached a level almost identical to that seen in early girdled plots (Fig. 1a).

Two factors will have contributed to the more rapid response of soil respiration in the late girdling treatment. First, temperate forest conifers preferentially allocate C to support shoot, needle and bud expansion in the early part of their growing season[10]. Second, at the same time roots can contain twice the concentration of starch seen later in the summer[11], as confirmed by our data from control plots (Table 1b). As a consequence, root respiration is less dependent on C import in the early growing season than in the late growing season.

On the basis of root biomass data from this forest stand[12], and the depletion of starch observed after the early girdling (Table 1b), we calculated an average loss of about 10 mg CO_2–C m^{-2} h^{-1}, which we assumed was due to use of starch during the first month after the early girdling. This corresponds to 10–20% of the respiratory flux from the soil during this period of June–July (Fig. 1a). The use of starch reserves after July in the early girdling treatment and by late girdled trees was calculated to be of less significance for the soil CO_2 efflux. Thus late, as opposed to early, girdling curtails C supply to the roots at a time when both flow of, and demand for, photosynthates are at a maximum.

With the respiration from early girdled plots used as a proxy for heterotrophic respiration, we calculated the levels of respiration by roots and their mycorrhizas by subtracting the values of respiratory activity on the early girdled plots from those on the control plots (Fig. 1b). Early in the summer, calculated root respiration increased later than heterotrophic respiration, whereas during the autumnal decline (Fig. 1c) heterotrophic and root respiration were similar (Fig. 1b). However, the apparently greater heterotrophic respiration early in the summer is probably an overestimate, owing to the 10 mg CO_2–C m^{-2} h^{-1} use of starch by the roots (see above) in the period immediately after girdling. When this amount is subtracted from the heterotrophic component, and, as is appropriate, added to the values obtained for root respiration, the curves for CO_2 production by the two compartments run closer together.

Several factors suggest that our estimates of a 52–56% contribution of root-mycorrhizal respiration to overall soil CO_2 efflux are conservative. Girdling can be expected not only to lead to a transient increase in the use of starch reserves of the roots, as observed by us (Table 1b), but also potentially to enhanced decomposition of starved roots and their symbionts by heterotrophs. Both of these processes should contribute to an increase, rather than a decrease, in soil respiration. In addition, all our plots supported small populations of the dwarf shrubs *Vaccinium vitis-idaea* L. and *Calluna vulgaris* L., whose root respiration will have contributed to the values attributed to non-root respiration from girdled plots. But even our conservative estimates of the contribution from root respiration are still considerably greater than those reported from the boreal zone, which were based on less direct and more intrusive methods.

In a study of *Picea mariana* in Canada, tree root respiration was calculated, on the basis of plant below-ground C allocation and estimates of root growth, to constitute only 24% of total soil respiration[3]. In a young Swedish Scots pine forest located 500 km to the south of the one used in our study, root respiration, also calculated on this basis, accounted for only 10 of the 63% of assimilates allocated below ground[13]. Our results contrast markedly with those of the only reported girdling experiment[14], which was carried out in temperate deciduous forest but found no effect of the treatment on soil respiration over the next 2 years. However, this experiment, which was unreplicated, used plots so small as to make edge effects likely.

Our experiment shows clearly that the flux of photosynthates has a big impact on soil respiratory losses—these being more influenced by seasonality than by variations in soil temperature. Thus, for example, the highest levels of root and total respiration were found in mid-August, well after both the maximum solar irradiance at the end of June and a long period of high air and soil temperatures in July (Fig. 1c). Previous studies have focused on the role of soil temperature as determinant of soil respiration (such as ref. 15). Our results indicate that the seasonal pattern of below-ground C allocation may be more important than soil temperature in determining root respiration. Improved models of forest C balances will be necessary to mitigate the process of global climate change, and we have shown that such models, when applied to the largest terrestrial ecosystem of the northern hemisphere, must take into account the fact that roots and their fungal symbionts contribute considerably more to soil respiratory C loss than acknowledged previously. Because these losses are driven directly by current assimilate allocated from the canopy, the coupling of photosynthetic activity to soil C efflux emerges as a key determinant of the ecosystem C budget, as has been suggested by a study of a temperate grassland[16]. Clearly, seasonal variations in these C allocation fluxes need to be considered in addition to the conventional physical parameters, if realism is to be achieved in assessing global C budgets. □

Methods

Girdling experiment

Nine square plots of 900 m^2 and 120 ± 12 trees each were established in a naturally regenerated 45–55-year-old Scots pine (*Pinus sylvestris* L.) forest at Åheden, northern Sweden (64° 14′ N, 19° 46′ E, 175 m above sea level). The soil is a weakly podzolized sediment of sandy silt. There is a sparse understorey of the dwarf shrubs *Vaccinium vitis-*

© 2001 Macmillan Magazines Ltd

idaea L. and *Calluna vulgaris* L. In early June 2000, the trees on three plots were girdled at 1.5-m above ground, by complete removal of the bark over 0.3-m long sections around the circumferences of the stems (early girdled plots). In mid-August 2000, we girdled the trees on three other plots (late girdled plots). On both occasions, it took 2 people 4 days to girdle all (~360) the trees. Three plots of ungirdled trees were used as controls throughout the experiment. Air temperature, soil temperature at 5-cm soil depth in the mineral soil, and precipitation were measured continuously at a standard meteorological station in an open field 150 m from the centre of the experiment (Fig. 1c). On one early girdled, one late girdled and one control plot, sensors below the organic mor layer (4-cm deep) recorded soil temperatures continuously (Fig. 1c). This, and more detailed measurements of surface soil temperatures and gravimetric soil moisture contents revealed no changes in response to the girdling treatments during the period of study.

Analysis of soil respiration

Soil respiration was measured as described previously[17] on nine occasions during June to October. Within 1 m of the centre of each plot, three 0.0464-m^2 plastic cylinders were placed on the ground before measurements, the superficial layer of lichens, mosses and litter was temporarily removed, and a lid was placed on each cylinder to form a 6-l head-space. Starting 2 min after the lid was put on, five 12-ml gas samples were removed by a syringe through a rubber membrane in the lid at 2-min intervals. At the two last sampling times, when soil respiration rates were low, the intervals between samples was 4 min. Gas samples were transferred to pre-evacuated vials, which were analysed for their contents of CO_2 on a gas purification module coupled on-line to an isotope-ratio mass spectrometer[17]. The rate of CO_2 evolution from the soil was calculated by linear regression. In mid-August, 2–3 days after the last girdling and during the optimal period of fruiting, sporocarps of fungi were collected from the ground in the central 100-m^2 area of each plot. These were identified by species, and classified into two functional categories, ectomycorrhizal or saprotrophic. The number of species, the number of sporocarps, and their total dry biomass was determined. We sampled roots for starch analysis on three occasions (Table 1b). Fine (< 2 mm diameter) roots of *P. sylvestris* were extracted from auger samples taken just outside the central 100-m^2 plots, dried (70 °C, 48 h) and analysed for starch[18]. For comparisons among treatments, we used plot mean values. In calculations of the contribution of root respiration to total soil respiration, we used treatment mean values.

Received 1 December 2000; accepted 29 March 2001.

1. Grace, J. & Rayment, M. Respiration in the balance. *Nature* **404,** 819–820 (2000).
2. Hanson, P. J., Edwards, N. T., Garten, C. T. & Andrews, J. A. Separating root and soil microbial contributions to soil respiration: a review of methods and observations. *Biogeochemistry* **48,** 115–146 (2000).
3. Malhi, Y., Baldocchi, D. D. & Jarvis, P. G. The carbon balance of tropical, temperate and boreal forests. *Plant Cell Environ.* **22,** 715–740 (1999).
4. Valentini, R. *et al.* Respiration as the main determinant of carbon balance in European forests. *Nature* **404,** 861–865 (2000).
5. Goulden, M. L. *et al.* Sensitivity of boreal forest carbon balance to soil thaw. *Science* **279,** 214–216 (1998).
6. Lindroth, A., Grelle, A. & Morén, A. S. Long-term measurements of boreal forest carbon balance reveal large temperature sensitivity. *Global Change Biol.* **4,** 443–450 (1998).
7. Post, W. M., Emanuel, W. R., Zinke, P. J. & Stangenberger, A. G. Soil carbon pools and world life zones. *Nature* **298,** 156–159 (1982).
8. Smith, S. E. & Read, D. J. *Mycorrhizal Symbioses* 2nd edn (Academic, San Diego, 1997).
9. Söderström, B. & Read, D. J. Respiratory activity of intact and excised ectomycorrhizal mycelial systems growing in unsterile soil. *Soil Biol. Biochem.* **19,** 231–236 (1987).
10. Hansen, J., Türk, R., Vogg, G., Heim, R. & Beck, E. in *Trees—Contributions to Modern Tree Physiology* (eds Rennenberg, H., Eschrich, W. & Ziegler, H.) 97–108 (Backhuys, Leiden, 1997).
11. Ericsson, A. & Persson, H. Seasonal changes in starch reserves and growth of fine roots of 20-year-old Scots pines. *Ecol. Bull.* **32,** 239–250 (1980).
12. Plamboeck, A. H., Grip, H. & Nygren, U. A hydrological tracer study of water uptake depth in a Scots pine forest under two different water regimes. *Oecologia* **119,** 452–460 (1999).
13. Ågren, G. I. *et al.* Annual carbon budget for a young Scots pine. *Ecol. Bull.* **32,** 307–313 (1980).
14. Edwards, N. T. & Ross-Todd, B. M. The effects of stem girdling on biogeochemical cycles within a mixed deciduous forest in eastern Tennessee. I. Soil solution chemistry, soil respiration, litterfall and root biomass studies. *Oecologia* **40,** 247–257 (1979).
15. Boone, R. D., Nadelhoffer, K. J., Canary, J. D. & Kaye, J. P. Roots exert a strong influence on the temperature sensitivity of soil respiration. *Nature* **396,** 570–572 (1998).
16. Craine, J. M., Wedin, D. A. & Chapin, F. S. III. Predominance of ecophysiological controls on soil CO_2 flux in a Minnesota grassland. *Plant Soil* **207,** 77–86 (1999).
17. Högberg, P. & Ekblad, A. Substrate-induced respiration measured *in situ* in a C_3-plant ecosystem using additions of C_4-sucrose. *Soil Biol. Biochem.* **28,** 1131–1138 (1996).
18. Steen, E. & Larsson, K. Carbohydrates in roots and rhizomes of perennial grasses. *New Phytol.* **104,** 339–346 (1986).

Acknowledgements

We gratefully acknowledge funding by the EU to P.H. and D.J.R. through the project FORCAST, by the Swedish Natural Sciences Research Council and Swedish National Energy Administration to P.H. and by the European Science Foundation to N.B. (LINKECOL); chemical analyses by A. Ohlsson and L. Skoglund; and technical assistance by G. Moen and L. Ohlsson.

Correspondence and requests for materials should be addressed to P.H. (e-mail: Peter.Hogberg@sek.slu.se).

Proximity signal and shade avoidance differences between early and late successional trees

Ian R. Gilbert*, Paul G. Jarvis† & Harry Smith‡

* *Institute of Cell and Molecular Biology; and † Institute of Ecology and Resource Management, University of Edinburgh, Edinburgh, EH9 3JU, UK*
‡ *Division of Plant Science, University of Nottingham, Sutton Bonington Campus, Loughborough LE12 5RD, UK*

Competitive interactions between plants determine the success of individuals and species. In developing forests, competition for light is the predominant factor. Shade tolerators acclimate photosynthetically to low light[1–3] and are capable of long-term survival under the shade cast by others, whereas shade avoiders rapidly dominate gaps but are overtaken in due course by shade-tolerant, later successional species. Shade avoidance[4–6] results from the phytochrome-mediated perception of far-red radiation (700–800 nm) scattered from the leaves of neighbours, provides early warning of shading[7], and induces developmental responses that, when successful, result in the overgrowth of those neighbours[8]. Shade tolerators cast a deep shade, whereas less-tolerant species cast light shade[9], and saplings tend to have high survivorship in shade cast by conspecific adults, but high rates of mortality when shaded by more-tolerant species[9]. Here we report a parallel relationship in which the shade-avoidance responses of three tree species are inversely proportional to proximity signals generated by those species. On this basis, early successional species generate small proximity signals but react strongly to them, whereas late successional species react weakly but generate strong signals.

Shade-avoiding plants respond to the relative amounts of red (R; 600–700 nm) and far-red (FR; 700–800 nm) photons in incident radiation by establishing different concentrations of the active 'Pfr' form of the phytochromes[10]. For many plants, the elongation growth rate of the stems bears an inverse linear relationship to the established concentration of Pfr, which is usually expressed as the proportion of total phytochrome (P) that is present as Pfr (that is, the photoequilibrium or Pfr/P)[11,12]. The slope of this relationship may be regarded as the 'shade avoidance response sensitivity'. This varies between species, with strong shade avoiders presenting a steep slope and weak shade avoiders presenting a shallow slope[13]. In addition, different species have different intrinsic elongation growth rates. These elements—intrinsic elongation rate and response sensitivity—combine to describe the overall shade avoidance response of plants in the natural environment. To explain the dynamics of tree growth in developing canopies, we sought to identify and quantify both intrinsic growth rates (R_i) and response sensitivities (S) of tree species with contrasting growth dynamics.

We grew stands of tree species at different spacings, and measured height growth, leaf area, leaf distribution and within-canopy radiation environment throughout several growing seasons. Three species of contrasting growth habits and leaf morphology were studied—*Acer pseudoplatanus* (a late successional species known as sycamore in the UK and sycamore-maple in North America), *Betula pendula* (silver birch, an early successional species) and *Populus deltoides* × *trichocarpa* cv. Beaupré (a hybrid poplar bred for rapid elongation growth). These species were grown in both pure stands (that is, single species) or in mixed stands that included another species, *Salix viminalis* (osier willow), chosen because it is a shrubby species with dense leaf coverage of stems and thereby

© 2001 Macmillan Magazines Ltd

generates strong proximity signals. The stands were rectangular grid arrays of differing spacings. Replicated monitoring over several years provided mean growth data from which the response parameters were derived. Radiation measurement was centred on selected 'target trees' and consisted of spectroradiometric measurements of R and FR radiation propagated roughly horizontally at different heights in the stands[12].

Height growth rate of the three species grown in pure stands increased as tree density increased (Fig. 1a), as did effects on the spectral distribution of the radiation that was propagated near-horizontally in the canopy, with a consequent depression of the R:FR ratio and thus of the calculated Pfr/P (Fig. 1b). All species reacted to neighbours by increased elongation growth, but the three species had different intrinsic growth rates. The slopes of the relationships between height growth rate and plant density, however, were not markedly different. The three species had differential effects on the radiation environment, with *A. pseudoplatanus* depressing Pfr/P the most and *B. pendula* the least. In other words, the late successional species generated strong proximity signals, whereas the early successional species generated weak signals.

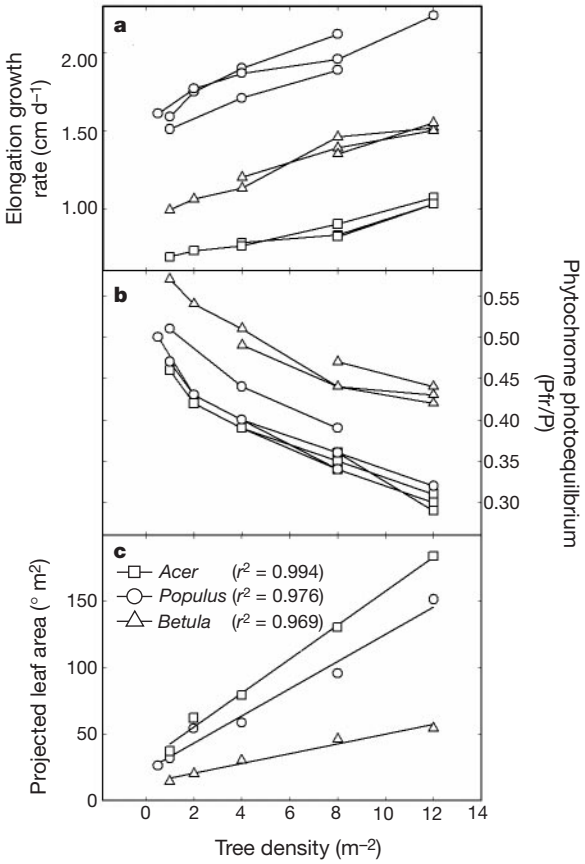


Figure 1 Relationships between planting density and tree height growth, within-canopy radiation environment and intensity of proximity signals. Pure stands of the three species were grown in regular grids at the stated planting densities. The radiation propagated horizontally at the central target tree was monitored with a spectroradiometer and the data converted to a value of expected phytochrome photoequilibrium (Pfr/P, where Pfr represents the amount of Pfr present at photoequilibrium and P represents the total phytochrome). **a**, Mean height growth over 120 days, from day 140 (20 May) to day 260 (17 September). **b**, Minimum photoequilibria recorded at day 220 (8 August), for each of three growing seasons. **c**, Intensity of reflection signals is a species-specific function of planting density. Mean data for projected leaf area were calculated from detailed measurements of four representative trees at each planting density. Lines are simple regression lines.

We obtained linear relationships when the tree data were replotted against Pfr/P (Fig. 2). The relationships include data from the final year of experimentation, in which we grew mixed plots at the two highest tree densities (black symbols). For *A. pseudoplatanus* and *Populus* cv. Beaupré, the mixed plot data lie within the ranges of Pfr/P obtained in pure stands of those species. By contrast, for *B. pendula* the mixed plot data show that the relationship extends beyond that obtained with the pure stands, which had much smaller effects on the R:FR ratio. In other words, *B. pendula* trees have stronger shade avoidance responses in stands with other species than in conspecific stands.

Figure 2 shows that the state of phytochrome regulates elongation growth in the three species in a very similar manner; thus, the elongation response of the target trees is determined solely by the effect of incoming radiation on the concentration of Pfr. Even so, the shade avoidance reactions of the three species in the plots were quite different, and consequently other factors, related to, or consequent on, the radiation signal must also determine the response. We reasoned that the element in the canopies that generates the radiative signal is likely to be the quantity of plant material, particularly the surface area of the leaves, that scatters FR radiation within the range of perception of the target tree, thereby generating proximity signals. Consequently, we calculated an effective radiation scattering quantity—called the 'projected leaf area' (PLA)—in each canopy established in these experiments. The calculation is facilitated by the regular grid pattern of the trees and by the limitation of each tree to a single main stem. The elements of the calculation are number of leaves, vertical distribution of leaf area, spatial distribution of leaf inclination, azimuth and zenith angles, and effective tree width and height (see Methods). We calculated PLA for the different species and densities using detailed measurements of four representative trees selected randomly in each plot, using a target tree at the centre of each stand.

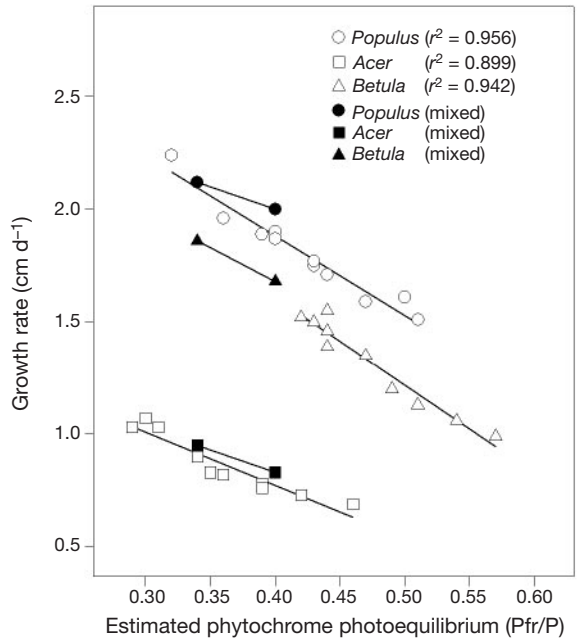


Figure 2 Height growth rate is a direct function of estimated phytochrome photoequilibrium. Data for the pure stands of single species (open symbols) and the mixed stands of all four species (closed symbols) are presented to show the linear relationships between phytochrome photoequilibrium estimated from the within-canopy radiation environment and tree height growth rate. Lines shown for the pure stands are regression lines for all annual data for each species; mixed plots at 8 and 12 trees per m^2 are separate lines connecting the two data items in each case.

© 2001 Macmillan Magazines Ltd

Figure 1c shows the relationships between tree density and PLA for the three species. In each case, there was an increase in PLA with increasing stand density of trees. The largest values of PLA were observed with *A. pseudoplatanus* and the smallest with *B. pendula*. This relates well to the large leaves of the former, which are densely distributed along the main axis (35 cm^2 of leaf area per cm of stem length), in contrast to the small leaves of the latter, which are scattered more thinly throughout the tree canopy (7 cm^2 of leaf area per cm of stem length).

Figure 3 confirms that PLA calculated as above is related to the strength of the proximity signal. We plotted PLA against Pfr/P estimated from spectroradiometric measurements of the radiation reaching the target trees. Over a very wide range, including both the pure stands and the mixed plots of all four species, Pfr/P was linearly related to the logarithm of PLA. These data relate the radiation signals received by the target tree to those generated by its neighbours. We draw the conclusion that leaves of the different species, irrespective of their detailed optical properties, have essentially the same effect on the R:FR radiation environment. On this basis, PLA provides a measure of the proximity signal generated by neighbours and received by a target tree in a stand of trees. Projected leaf area is therefore an effective estimate of the capacity of a tree to generate radiation signals as a function of leaf area density. We generated Fig. 4 by plotting elongation growth rate against PLA, using all the available data. This depicts the shade avoidance capacity of the three species by relating the scale of response to the source of the signal—that is, to the amount of leaf material scattering radiation in the perception range of the target tree.

Using Fig. 4, we calculated R_i from the *y*-axis intercepts and *S* from the slopes of the lines. Table 1 shows that *A. pseudoplatanus* has a low intrinsic elongation rate and a small response sensitivity to proximity signals. *A. pseudoplatanus* is generally regarded as a late successional tree, capable of substantial shade tolerance. Eleven species of maple in North American forests were shown to be shade tolerant and did not show shade avoidance responses in gap-edge situations[14]. By contrast, *B. pendula* is an early successional species, with a low intrinsic elongation rate but a very high capacity to respond to radiation signals. On this basis, *B. pendula* is more than five times as responsive as *A. pseudoplatanus*. The hybrid *Populus* cultivar Beaupré, bred for high productivity, has a high intrinsic growth rate but shade avoidance sensitivity intermediate between that of *A. pseudoplatanus* and *B. pendula*.

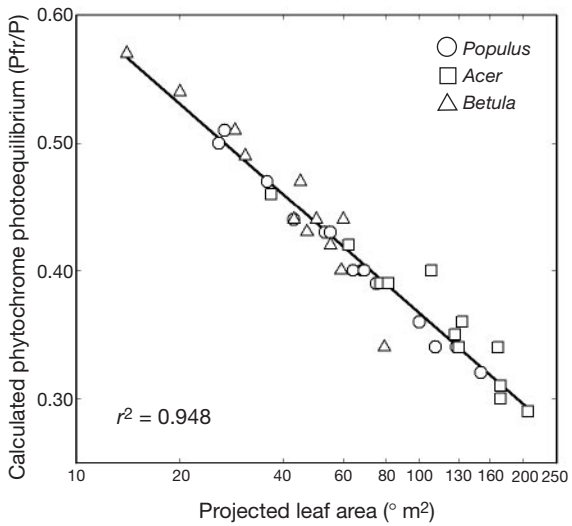


Figure 3 Projected leaf area is an accurate estimate of proximity signals. Estimated Pfr/P ratio established by the radiation reaching the target tree is shown to bear a close logarithmic relationship to projected leaf area.

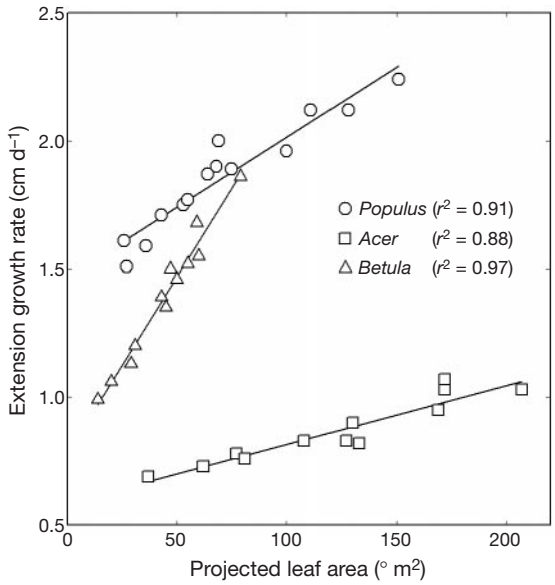


Figure 4 Projected leaf area has a species-specific relationship to the height growth rates of the three species. Data drawn from all annual trials, including both pure and mixed plots.

Shade avoidance has major fitness attributes in crowded communities, but may be disadvantageous for some species in that it results in the allocation of resources to stem growth[15] and away from storage and reproductive structures. For late successional trees such as *A. pseudoplatanus*, the large leaves and dense canopy effectively generate strong FR proximity signals. Under the circumstances of these experiments, *A. pseudoplatanus*, which is generally regarded as shade tolerant, showed shade avoidance responses over a wide range of PLA, but the sensitivity of response to such signals was low. *B. pendula*, an early successional tree, has sparsely distributed small leaves and generates weak proximity signals, although it responds strongly to such signals. If FR proximity signals are indicators of impending competition for radiant energy[7], then *B. pendula* has the capacity to respond well to competition but generates only small radiation signals itself. Thus, *B. pendula* seedlings appearing in an established *A. pseudoplatanus* woodland would experience strong proximity signals and exhibit shade avoidance responses that would be maladaptive. In a gap, however, emerging *B. pendula* seedlings, with a higher intrinsic elongation rate than that of *A. pseudoplatanus* seedlings and with a strong response to even weak proximity signals, would be well-equipped to colonize the gap, at least temporarily. Ultimately, the shade-tolerance of *A. pseudoplatanus*, plus its parsimonious intrinsic elongation rate, would allow it to survive under the light shade cast by *B. pendula*.

We conclude that the morphology and size of leaves, and their distribution in relation to height in the canopy, are highly significant in the competition for living space that exists at all times

Table 1 Intrinsic elongation growth rates and shade avoidance response sensitivities

Species	Intrinsic elongation rate (R_i)	Shade avoidance sensitivity (*S*)	Relative sensitivity
Acer	0.58	0.0024	1
Betula	0.78	0.0138	5.75
Populus	1.47	0.0055	2.29

Intrinsic elongation growth rates (cm d^{-1}) were calculated from the intercepts of the regression lines in Fig. 4 with a value of zero for RLA. Shade avoidance response sensitivities were calculated from the slopes of the respective regression lines in Fig. 4b.

© 2001 Macmillan Magazines Ltd

between plants growing in crowded communities—both in terms of generating proximity signals and in terms of the capacity for shade tolerance. These data provoke speculation that there may be an evolutionary trade-off between the morphology, architecture and growth pattern of a shade tolerator and those of a shade avoider. □

Methods

Planting materials and growth of canopies

This research was carried out at the University of Leicester, UK. *Betula pendula* saplings were obtained, with permission, from a reserve run by the Royal Society for the Protection of Birds at Brampton, Northumbria, UK (grid reference NY587655). *A. pseudoplatanus* saplings were obtained, with permission, from Priory Wood—a designated 'site of special scientific interest' (SSSI) in Leicestershire, UK (grid reference SK498109). The *S. viminalis* plants were grown from stem cuttings taken originally (1971) from a clonal coppice stand in south Nottinghamshire, UK, by H.S. Hybrid *P. deltoides* × *trichocarpa* cv. Beaupré were from the Institute of Terrestrial Ecology, Penicuik, UK. Trees were grown in 10-dm^3 pots irrigated individually, and top dressed in early spring and summer with 10 g of Osmocote Plus (15:11:13) and 5 g of Greenmaster Super-N (24:0:0). We arranged pots in stands of 1, 2, 4, 8 and 12 pots per m^2 either on the roof of the Biological Sciences building or at the University of Leicester Botanic Gardens, and grew them for at least three seasons. All trees were pruned as they grew to remove laterals to keep the three-dimensional geometry of the plants simple and to minimize azimuthal effects. Data for tree height, tree width, numbers of leaves and leaf area were recorded throughout the growing seasons. Leaf areas were estimated from a programme of detailed measurements in mid-season near day 220 (8 August). All the above variables were measured for eight trees in each plot, except for the detailed leaf measurements, for which only four trees were measured.

Radiation environment

For all canopies, a central 'target tree' was selected, and the radiation that propagated more or less horizontally within the canopy and reached the tree was monitored throughout the seasons. Horizontal narrow angle sensors (10° × 180° letterbox), placed at six pre-selected height positions at the target tree, piped radiation along fibre-optic bundles to a selector wheel and then to a spectroradiometer (Licor LI-1800; Lincoln, NV, USA). We collected data every 10 min, giving a complete set of readings for the six height positions during daylight hours[12]. Spectral photon flux data were converted into estimated phytochrome photoequilibrium (Pfr/P) by integrating the absorption cross-sections of Pr and Pfr with the spectral photon flux distributions of the incident radiation across the 400–800-nm waveband[16].

Calculation of projected leaf area

Projected leaf area (PLA) is an estimate of the total leaf area in the neighbourhood of the target tree that is capable of scattering radiation towards the target tree. To estimate PLA, we measured leaf position (that is, height above ground), leaf area and leaf angles (azimuth, inclination and zenith angles) for four randomly selected signal trees in each plot, and obtained estimates of the means from these. For each leaf distributed along a tree axis, the height and angle data were used to calculate the area projected by the leaf onto a cylinder representing the tree surface, itself calculated from tree height and average crown diameter. The total of the projected areas for all leaves on the tree was divided by the total area of the cylinder surface, and this ratio was used to calculate occlusion of the signal by intervening trees. The occlusion of the signal from distant trees by nearer trees was calculated geometrically using the grid of the planting design. The reduction of signal with tree distance was calculated as a function of the angular wedge subtended by a tree cylinder to the stem of the target tree. PLA represents the arithmetic sum of all signals, corrected for occlusion and for distance. The units of PLA are degrees × area (° cm^2).

Received 22 February; accepted 3 April 2001.

1. Boardman, N. K. Comparative photosynthesis of sun and shade plants. *Annu. Rev. Plant Physiol.* **28,** 355–377 (1977).
2. Bjørkman, O. in *Encyclopaedia of Plant Physiology* Vol. 12.4 (eds Lange O. L., Nobel, P. S., Osmond, C. B. & Ziegler, H.) 57–107 (Springer, Berlin, 1981).
3. Kull, O & Niinemets, U. Distribution of leaf photosynthetic properties in tree canopies: comparison of species with different shade tolerance. *Funct. Ecol.* **12,** 472–479 (1998).
4. Smith, H. Light quality, photoperception and plant strategy. *Annu. Rev. Plant Physiol.* **33,** 481–518 (1982).
5. Smith, H. Physiological and ecological function within the phytochrome family. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **46,** 289–315 (1995).
6. Smith, H. Phytochromes and light signal perception by plants—an emerging synthesis. *Nature* **407,** 585–591 (2000).
7. Ballare, C. L., Scopel, A. L. & Sanchez, R. A. Far-red radiation reflected from adjacent leaves: an early signal of competition in plant canopies. *Science* **247,** 329–332 (1990).
8. Casal, J. J. & Smith, H. The function, action and adaptive significance of phytochrome in light-grown plants. *Plant Cell Environ.* **12,** 855–862 (1989).
9. Canham, C. D., Finzi, A. C., Pacala, S. W. & Burbank, D. H. Causes and consequences of resource heterogeneity in forests—interspecific variation in light transmission by canopy trees. *Can. J. Forest Res.* **24,** 337–349 (1994).
10. Smith, H. & Holmes, M. G. The function of phytochrome in the natural environment. III. Measurement and calculation of phytochrome photoequilibrium. *Photochem. Photobiol.* **25,** 547–550 (1977).
11. Morgan, D. C. & Smith, H. Linear relationship between phytochrome photoequilibrium and growth in plants under simulated natural radiation. *Nature* **262,** 210–212 (1976).
12. Gilbert, I. R., Seavers, G. P., Jarvis, P. G. & Smith, H. Photomorphogenesis and canopy dynamics. Phytochrome-mediated proximity perception accounts for the growth dynamics of canopies of *Populus trichocarpa* × *deltoides* 'Beaupré'. *Plant Cell Environ.* **18,** 475–449 (1995).
13. Morgan, D. C., Smith, H. A systematic relationship between phytochrome-controlled development and species habitat for plants grown in simulated natural radiation. *Planta* **145,** 253–259 (1979).
14. Lei, T. T & Lechowicz, M. J. Diverse responses of maple saplings to forest light regimes. *Am. J. Bot.* **82,** 9–19 (1998).
15. Child, R., Morgan, D. C. & Smith, H. Control of development in *Chenopodium album* by shadelight: The effect of light quality (red:far-red ratio) on morphogenesis. *New Phytol.* **89,** 545–555 (1981).
16. Hayward, P. M. in *Techniques in Photomorphogenesis* (eds Smith, H. & Holmes, M. G.) 159–173 (Academic, London, 1984).

Acknowledgements

This research was funded by the Natural Environment Research Council, UK, and was made possible by the technical assistance of S. Smith, M. Pratt and G. Benskin.

Correspondence and requests for materials should be addressed to H.S. (e-mail: Harry.Smith@Nottingham.ac.uk).

Ant odometry in the third dimension

Sandra Wohlgemuth*, Bernhard Ronacher* & Rüdiger Wehner†

* *Institute of Biology, Humboldt-University Berlin, Invalidenstrasse 43, D 10099 Berlin, Germany*
† *Department of Zoology, University of Zurich, Winterthurerstrasse 190, CH 8057 Zurich, Switzerland*

Desert ants (*Cataglyphis*) are renowned for their ability to perform large-scale foraging excursions and then return to the nest by path integration. They do so by integrating courses steered and the distances travelled into a continually updated home vector[1]. Whereas the angular orientation is based on skylight cues[2], how the ants gauge the distances travelled has remained largely unclear[3,4]. Furthermore, almost all studies on path integration in *Cataglyphis*[5,6], as well as in spiders[7,8], rodents[9], and humans[10,11], have aimed at understanding how the animals compute home-bound courses in the horizontal plane. Here, we investigate for the first time how an animal's odometer operates when a path integration task has to be accomplished that includes a vertical component. We trained *Cataglyphis* ants within arrays of uphill and downhill channels, and later tested them on flat terrain, or vice versa. In all these cases, the ants indicated homing distances that corresponded not to the distances actually travelled but to the ground distances; that is, to the sum of the horizontal projections of the uphill and downhill segments of the ants' paths.

Saharan desert ants (*Cataglyphis fortis*) leave their underground colonies for distances of several hundred metres, winding their way in a tortuous search for food, but then return to their point of departure along an amazingly straight path[5]. They accomplish this task by integrating information about the angular and linear components of movement, information provided by a compass and an odometer. The compass, which exploits polarization and spectral gradients in the sky, has been studied in some behavioural and neurophysiological detail[2,12], but attempts to unravel the ant's odometer have so far met with limited success. Self-induced visual flow might be used to some extent[3], as in flying bees, in which the odometer depends mainly on flow-field information[13–15]. Walking desert ants, however, can do without it and can derive information about walking distances completely from proprioceptive inputs[3,4]. Here, we investigate how these ants measure travelling distances in undulating terrain.

© 2001 Macmillan Magazines Ltd

In a first experiment, we trained ants to walk over a series of hills, experimentally mimicked by a linear array of (in this case nine) uphill and downhill channels (slopes +54° and −54°, respectively), until they reached a feeding site. The distance travelled by the ants within these tilted channels was 8.7 m, which corresponded to a ground distance of 5.2 m. When later tested within a horizontal channel (slope 0°), the ants taken from the feeder walked for a distance of 4.7 ± 1.0 m (Fig. 1a). Then they made a sharp U-turn (180°) and started long-lasting oscillating search movements about that point. That is, the distance travelled by the ants in the flat channel was much shorter than the one travelled by them during their preceding outbound runs across the hills ($P \ll 0.001$ (Wilcoxon)) but corresponded approximately to the ground distance between nest and feeder. This finding is corroborated by the reciprocal experiment, in which the ants were trained to walk for 5.2 m within the flat channel and were later tested in the hilly array (Fig. 1b). There, they covered an actual travel distance of 8.1 ± 2.2 m, which was much larger than the one covered by the same animals in the flat training control channel ($P < 0.001$ (Wilcoxon)). The corresponding ground distance (4.8 ± 1.3 m) is, however, not significantly different from the control ($P > 0.1$ (Wilcoxon)). Obviously, when the ants walk along an uphill and downhill path, their odometer records the ground distance rather than the distance actually travelled.

We further tested whether the projection rule also works for other three-dimensional profiles. Owing to the geometrical layout of the channel array used in the first experiment, the cumulative ascent walking distances were 4.5 m (cumulative descents were 4.2 m) and thus close to the means observed in the tests described above (open bars in Fig. 1a, b). Thus, it might have been only the ascent (or only the descent) walking distances on which the ants had relied. To resolve this ambiguity, we trained ants over seven hill segments within the symmetric array (Fig. 2), but tested them within asymmetric arrays consisting of series of steep ascents (+54°, 50 cm) and shallow descents (−24°, 100 cm), or vice versa (Fig. 2). Let us assume, for instance, that *Cataglyphis* would switch on its odometer only during its uphill runs and that the odometer treated these uphill runs as though they were horizontal. Then, with the steep-ascent array, it should walk over a ground distance of 8.5 m (filled asterisk in Fig. 2) rather than the 4.2-m training distance. However, contrary to this assumption, the observed ground distances for the three test situations did not differ significantly from each other or from the 4.2-m ground distance experienced during training (Fig. 2; $P > 0.05$ (Friedman)). An analogous argument can be used to reject the assumption that the ants measure distances only during their downhill paths (Fig. 2; lower column). Similar results were obtained with animals that were also trained, rather than only tested, in an asymmetrical channel (five hills, +24° ascent, −54° descent as seen from the nest). Again, the (horizontal) search distances in the two reciprocally oriented asymmetrical test channels were not significantly different (6.12 m and 6.39 m, compared to the 6.15-m ground distance experienced in the training situation; $P > 0.1$). The conclusion to be drawn from these results (Figs 1 and 2) is that, whenever *Cataglyphis* ants walk over hilly terrain, their odometer records the ground distances rather than the distances actually travelled. Hence, the ants must be able to perceive the steepness of the uphill and downhill walking platforms, and to include this information into their distance estimation process.

Even though, at present, the mechanisms used by the ants in accomplishing this task are not known, the results presented here allow at least the following conclusions. Two cues that have been repeatedly invoked as means by which insects could gauge distances are the measurement of either energy expenditure or time spent walking[16,17]. Our results show that the mere energy expended by the animals on their way from the nest to the feeder (irrespective of other odometric cues) cannot be a useful means of gauging distances travelled. For example, in the hilly test array, the ants covered a distance 1.5 times greater than in the flat channel in which they had been trained (Fig. 1b). Even if one assumed that walking over hills does not cause additional costs, in the experimental

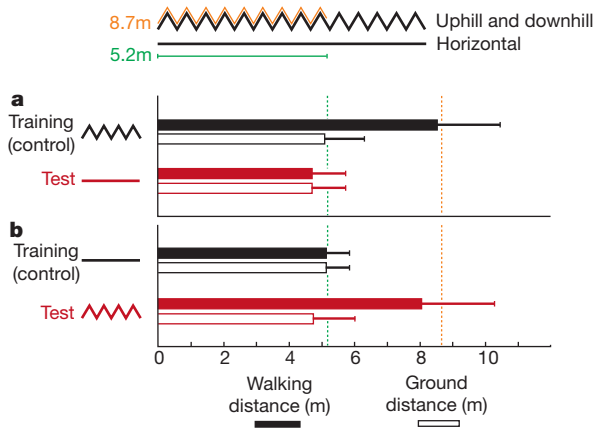


Figure 1 The ant's odometer does not record the distance actually travelled along an uphill–downhill path but rather the horizontal projection of that path (that is, the ground distance). **a**, Training along a symmetrical uphill-and-downhill path (a schematic side view is shown at the top). Ants were trained to walk over nine hills to a food source 8.7 m away and were then tested either in a control channel that had the same dimensions as the training channel (black) or in a horizontal (flat) test channel (red). Filled bars indicate actual walking distances covered in the test channels (means + s.d.), open bars indicate the corresponding ground distances (in the flat channel, the two values are identical). The green dotted line indicates the expected ground distance value and the orange dotted line indicates the expected walking distance value. In the training control, the walking and ground distances do not differ significantly from the expectations. In the horizontal test channel, the travelling distance was much shorter than the distances travelled in the hilly array (orange dotted line). Rather, the walking distance corresponded to the ground distance of the training control. Twenty-one animals were tested in both conditions. **b**, Training in a horizontal channel (to a food source 5.2 m away). The green dotted line indicates the expectation for ground and walking distance and the orange dotted line indicates the uphill–downhill walking distance that corresponds to a 5.2-m walking distance in the flat channel. In the hilly test channel, the walking distance is significantly greater than the control but the mean ground distance is not. This experiment involved 17 animals.

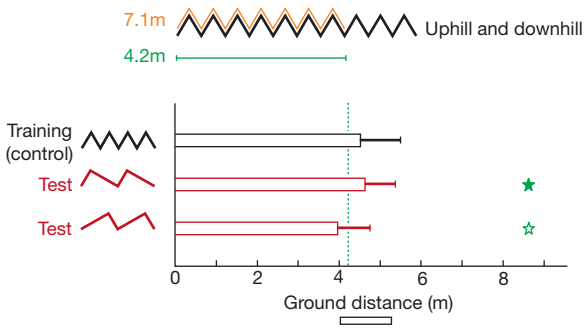


Figure 2 In asymmetrical arrays of uphill and downhill segments ants also use the ground distances. Tests within the asymmetrical uphill-and-downhill array, training along the symmetrical array (over seven symmetrical hills; 7.1 m walking distance, 4.2 m ground distance; inclinations of tests channels were +54° ascent and −24° descent, or +24° ascent and −54° descent). The mean ground distances covered by the ants within the three test arrays did not differ significantly. The green dotted line indicates the ground distance covered during training. The green asterisks indicate the expected ground distances covered under the assumption that the ants switch on their odometer only during the uphill (filled asterisk) or downhill (open asterisk) runs. This experiment involved 15 animals. This figure shows that the ants record neither the total distance travelled (compare with Fig. 1) nor only the cumulative ascent (or only the cumulative descent) distances travelled, but always the ground distances.

© 2001 Macmillan Magazines Ltd

situation mentioned above the ants experienced at least 1.5 times more energy consumption when returning over the hills than on flat terrain. By contrast, when trained over the hills and tested in the flat channel, the ants stopped at about 55% of the walking distance covered in the hilly array (Fig. 1a). Thus, if the ants relied on energy expenditure alone, they would have grossly overestimated their walking distance in one experimental arrangement or underestimated it in the other. Of course, if walking uphill and downhill were more costly for the ants than walking on level ground, this discrepancy would be even larger.

The ants' walking speed was much reduced in both the (+54°) uphill and (−54°) downhill channels ($6.1 \pm 2.3\ \mathrm{cm\,s^{-1}}$ and $9.8 \pm 4.7\ \mathrm{cm\,s^{-1}}$, respectively, compared with $17.0 \pm 2.9\ \mathrm{cm\,s^{-1}}$ in the flat channel). The observation that the walking speeds were reduced (compared with the horizontal) even when the ants walked downhill indicates that walking downhill does not necessarily allow a simple recovery of some of the energy spent during the uphill paths[18]. In conclusion, our results make energy expenditure without additional information about, for example, the slope of the walking platform an unlikely means of gauging distances. Instead, they accord with recent evidence obtained in *Cataglyphis*[19] and in flying bees[14,15,20] by quite different experimental procedures.

The same arguments apply to time as a potential odometric cue. Again, it cannot be time as an isolated parameter that provides the ants with the necessary information about distances travelled. In the first experiment, for instance, the ants spent four to five times longer when homing across the hills than when homing on a level ground (124 ± 49 s and 27.3 ± 7.6 s, respectively, for the data from Fig. 1a; 99 ± 39 s and 26.5 ± 7.9 s, respectively, for the data from Fig. 1b). An obvious way to cope with differences in walking speeds would be to monitor the frequency of a central pattern generator (CPG) for leg movements and to increase or decrease the total time spent walking in proportion to deviations from a basic CPG frequency. However, in analogy to the invoked energy cue, the data from Fig. 1 demonstrate that, under this assumption, the ants must have grossly misjudged velocity (or total time) in at least one of the experimental arrays used.

Finally, our experimental arrangement also rules out gauging distance by integrating self-induced optic flow. The floor and the walls of the channels were uniformly grey and hence devoid of any contrast cues that could have been exploited by the ants' visual system. Furthermore, and unlike what has recently been observed in flying bees[13–15,20], flow-field cues have only a negligible role in distance estimation in walking ants[3,4].

These results leave information derived from the animal's own movements (idiothetic cues[21]) as the most likely source of information used by the *Cataglyphis* odometer. In ants, gravity perception, and hence most probably also the measurement of the slopes of the walking platforms, is mediated by proprioreceptors located on the joints between ant's main body parts: between head and thorax, between thorax, petiole and gaster, and between thorax and the coxae of the legs[22]. The fact that a variety of hairplates contribute to gravity perception and might complement each other when surgically eliminated[22] makes it a tricky task to unravel the mechanisms by which the ants gauge ground distances with a high precision when walking over hilly terrain.

Whatever the mechanisms are by which this proprioceptive input is integrated to obtain information about slopes, our present results already show that, in gauging distances travelled, the ants must use more sophisticated means than merely counting numbers of steps or monitoring the output of a CPG. In more general terms, we can conclude that, while ants are foraging and homing across uneven terrain, they project their incremental path segments onto the horizontal and perform path integration, so to speak, within a virtual x–y plane. As the foraging ants cannot foresee while they are winding their way across hilly terrain what the three-dimensional structure of the area will be across which they have to return to the nest, performing path integration within a (virtual) horizontal x–y plane will substantially reduce navigational errors.

Next, we are going to investigate how *Cataglyphis* performs path integration in a fully three-dimensional set-up; that is, not only in the vertical x–y plane, but including the y dimension as well. In preliminary ways, this question has been addressed in spiders[23] and mice[24]. In both cases, however, the distance component has been excluded from the analyses, and only the animals' directional choices have been tested. Hence, how an animal accomplishes and uses vector integration in three-dimensional space is still an open question. □

Methods

Animals and study area

All experiments were performed on individually marked foragers of *C. fortis* in a salt pan near the village of Maharès, Tunisia (34.58° N, 10.50° E).

Experimental set-up

In a first experiment, the ants were trained from their nest to a food source (water melon and biscuit morsels) in a U-shaped aluminium channel (7 cm base, 7 cm walls), which forced the animals to walk over a series of nine symmetrical hills. Each hill segment had side lengths of 50 cm with a slope (to the horizontal) of 54° on both sides. During this training, the ants covered a walking distance of 8.7 m, which corresponded to a ground distance of 5.2 m. A forager arriving at the food source was taken in the middle of the last descent and transferred to perform its homebound run in one of two test channels, which were aligned in parallel to the training channel. One test channel consisted of 15 hill segments (same dimensions as in training; total ground distance 8.85 m; walking distance 15 m). The other test channel was horizontal (length 10.5 m, same cross-sectional dimensions). The bases of all channels were covered with glued sand, to facilitate the ant's mountaineering. The side walls were sprayed with matt grey paint to exclude optic flow cues and to avoid disturbing reflections. In a second experiment, a different group of ants experienced the reciprocal situation: these animals were trained in a horizontal channel (to a food source 5.2 m away) and then transferred to one of the two test channels (horizontal or over hills (Fig. 1b)). As a control, ants were trained in the symmetrical hill channel to a food source 4.2 m away (ground distance; corresponding to seven hill segments) and then tested in one of three channels: one channel consisting of symmetrical hills, and two channels with asymmetrical hill segments (Fig. 2). The asymmetric hill segments had side lengths of 0.5 m and 1.0 m, and slopes of 54° and 24°, respectively (1.21-m ground distance). Both orientations were used (steep ascent and shallow descent, and shallow ascent and steep descent).

Data collection and statistics

We recorded the point at which the ants switched from their steady straight path to their typical nest-searching behaviour[25]. This transition is marked by a very conspicuous 180° turn, followed by a backwards path[4]. In all figures, the average distances to these first U-turns are defined as the ants' searching distance. The travelling times for the uphill and downhill segments were recorded separately and used to calculate the ants' average speed during ascents and descents. The data in Figs 1 and 2 were obtained in a repeated measures design by testing individually marked ants in different channels. We therefore applied Wilcoxon's signed rank test to the data in Fig. 1 and Friedman's test combined with the Wilcoxon–Wilcox test to the data in Fig. 2. A group of 21 animals participated in the experiment in Fig. 1a, 17 different animals were tested in the experiment shown in Fig. 1b and another 15 animals were used in the asymmetrical channels (Fig. 2).

Received 21 December 2000; accepted 23 April 2001.

1. Müller, M. & Wehner, R. Path integration in desert ants, *Cataglyphis fortis*. *Proc. Natl Acad. Sci. USA* **85,** 5287–5290 (1988).
2. Wehner, R. in *Orientation and Communication in Arthropods* (ed. Lehrer, M.) 145–185 (Birkhäuser, Basel, 1997).
3. Ronacher, B. & Wehner, R. Desert ants *Cataglyphis fortis* use self-induced optic flow to measure distances travelled. *J. Comp. Physiol. A* **177,** 21–27 (1995).
4. Ronacher, B., Gallizzi, K., Wohlgemuth, S. & Wehner, R. Lateral optic flow does not influence distance estimation in the desert ant *Cataglyphis fortis*. *J. Exp. Biol.* **203,** 1113–1121 (2000).
5. Wehner, R. & Wehner, S. Insect navigation: use of maps or Ariadne's thread. *Ethol. Ecol. Evol.* **2,** 27–48 (1990).
6. Müller, M. & Wehner, R. The hidden spiral: systematic search and path integration in desert ants, *Cataglyphis fortis*. *J. Comp. Physiol. A* **175,** 525–530 (1994).
7. Seyfarth, E. A., Hergenröder, R., Ebbes, H. & Barth, F. G. Idiothetic orientation of a wandering spider: compensation of detours and estimates of goal distance. *Behav. Ecol. Sociobiol.* **11,** 139–148 (1982).
8. Görner, P. & Claas, B. in *Neurobiology of Arachnids* (ed. Barth, F. G.) 275–297 (Springer, Berlin, 1985).
9. Etienne, A. S., Maurer, R. & Seguinot, V. Path integration in mammals and its interaction with visual landmarks. *J. Exp. Biol.* **199,** 201–209 (1996).
10. Sauve, J. P. *L'orientation spatiale: Formalisation d'un modèle de mémorisation égocentrée et expérimentation chez l'homme.* PhD thesis, Univ. Aix-Marseille (1989).
11. Loomis, J. M. *et al.* Nonvisual navigation by blind and sighted: assessment of path integration ability. *J. Exp. Psychol. Gen.* **122,** 73–91 (1993).
12. Wehner, R. in *Neural basis of Behavioural Adaptation* (eds Schildberger, K. & Elsner, N.) 103–143 (G. Fischer, Stuttgart, 1994).

© 2001 Macmillan Magazines Ltd

13. Srinivasan, M. V., Zhang, S. W., Lehrer, M. & Collett, T. S. Honeybee navigation *en route* to the goal: visual flight control and odometry. *J. Exp. Biol.* **199,** 237–244 (1996).
14. Srinivasan, M. V., Zhang, S. W. & Bidwell, N. J. Visually mediated odometry in honeybees. *J. Exp. Biol.* **200,** 2513–2522 (1997).
15. Srinivasan, M. V., Zhang, S., Altwein, M. & Tautz, J. Honeybee navigation: nature and calibration of the 'odometer'. *Science* **287,** 851–853 (2000).
16. von Frisch, K. *The Dance Language and Orientation of Bees* (Harvard Univ. Press, Cambridge, Massachusetts, 1967).
17. Heran, H. Ein Beitrag zur Frage nach der Wahrnehmungsgrundlage der Entfernungsweisung der Bienen (*Apis mellifica* L.). *Z. Vergl. Physiol.* **38,** 168–218 (1956).
18. Taylor, C. R., Caldwell S. L. & Rowntree, V. J. Running up and down hills: some consequences of size. *Science* **178,** 1096–1097 (1972).
19. Schäfer, M. & Wehner, R. Loading does not affect measurement of walking distance in desert ants *Cataglyphis fortis*. *Verh. Deutsch. Zool. Ges.* **86,** 270 (1993).
20. Esch, H. E. & Burns, J. E. Distance estimation by foraging honeybees. *J. Exp. Biol.* **199,** 155–162 (1996).
21. Mittelstaedt, H. & Mittelstaedt, M. Mechanismen der Orientierung ohne richtende Aussenreize. *Fortschr. Zool.* **21,** 46–58 (1973).
22. Markl, H. Borstenfelder an den Gelenken als Schweresinnesorgane bei Ameisen und anderen Hymenopteren. *Z. Vergl. Physiol.* **45,** 475–569 (1962).
23. Hill, D. E. Orientation by jumping spiders of the genus *Phidippus* (Araneae: Salticidae) during the pursuit of prey. *Behav. Ecol. Sociobiol.* **5,** 301–322 (1979).
24. Bardunias, P. M. & Jander, R. Three dimensional path integration in the house mouse. *Naturwissenschaften* **87,** 532–534 (2000).
25. Wehner, R. & Srinivasan, M. V. Searching behaviour of desert ants, genus *Cataglyphis* (Formicidae, Hymenoptera). *J. Comp. Physiol.* **142,** 315–338 (1981).

Acknowledgements

We thank H. Gansner for her help in running the experiments and the members of the Zurich–Maharès crew for their co-operation in the field. We further thank H. Heise for the construction of the channel arrays and U. Menzi and H. Michel for their help in designing the figures and preparing the manuscript. Financial support came from the Swiss National Science foundation and the G. & A. Claraz foundations, grants to R.W..

Correspondence and requests for materials should be sent to B.R. (e-mail: bernhard.ronacher@rz.hu-berlin.de)

Motion-induced blindness in normal observers

Yoram S. Bonneh*, Alexander Cooperman† & Dov Sagi†

** Smith–Kettlewell Eye Research Institute, 2318 Fillmore Street, San Francisco, California 94115, USA*
† Department of Neurobiology, Brain Research, The Weizmann Institute of Science, Rehovot, 76100 Israel

Cases in which salient visual stimuli do not register consciously are known to occur in special conditions, such as the presentation of dissimilar stimuli to the two eyes[1] or when images are stabilized on the retina[2]. Here, we report a striking phenomenon of 'visual disappearance' observed with normal-sighted observers under natural conditions. When a global moving pattern is superimposed on high-contrast stationary or slowly moving stimuli, the latter disappear and reappear alternately for periods of several seconds. We show that this motion-induced blindness (MIB) phenomenon is unlikely to reflect retinal suppression, sensory masking or adaptation. The phenomenology observed includes perceptual grouping effects, object rivalry and visual field anisotropy. This is very similar to that found in other types of visual disappearance, as well as in clinical cases of attention deficits, in which partial invisibility might occur despite the primary visual areas being intact[3]. Disappearance might reflect a disruption of attentional processing, which shifts the system into a winner-takes-all mode, uncovering the dynamics of competition between object representations within the human visual system.

The phenomenon reported here adds to a class of known phenomena of 'visual disappearance' in which salient stimuli disappear from visual awareness, as if erased in front of the observer's eyes. Such phenomena, in which information is ignored owing not to a failure to notice[4] but to explicit 'erasing', include binocular[1,5] and monocular[6,7] rivalry (in which superimposed dissimilar patterns presented to different eyes or in different colours disappear in alternation), stabilized images that fade away[2], afterimages that similarly disappear and reappear[8], and Troxler fading (in which low-contrast peripheral stimuli disappear under strict fixation[9]). Clinical cases of explicit disappearance have also been reported in patients with simultanagnosia[10,11]. The phenomenon of MIB occurs in normal observers under normal (monocular) viewing conditions and might occur in natural situations. It was first described by Grindley and Townsend[12,13], who studied 'movement masking' in binocular rivalry, but went largely ignored until now, probably because its compelling strength in normal viewing was never observed.

We presented high-contrast yellow patterns (targets) together with a dynamic blue random dot pattern (mask), as described in Fig. 1. With steady fixation, but not with strict fixation (small eye movements could be tolerated), observers reported seeing long periods (several seconds) of complete disappearance of one or more target patterns, which disappeared and reappeared in a seemingly spontaneous way. We used the accumulated duration of disappearance as a measure for studying the parameters that affect this phenomenon and the mechanisms involved. Several hundred observers[14] have already confirmed the effect by subjective report, and very few have failed to experience MIB.

The results are summarized in Fig. 2. The properties of MIB do not seem to reflect sensory suppression or adaptation. First, targets of higher luminance contrast disappeared more (Fig. 2a), as opposed to the Troxler fading effect[15], and thus cannot be explained by a contrast-gain-control mechanism. Second, moving or dynamic targets disappeared too (Fig. 2c, d), producing the striking phe-

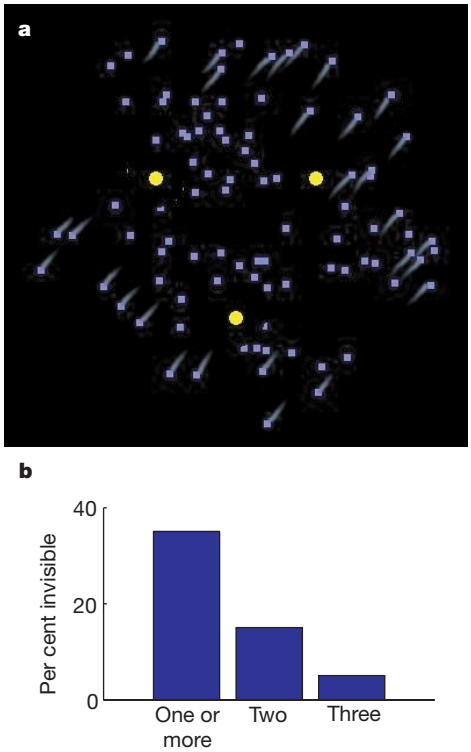


Figure 1 Typical stimuli used to measure the motion-induced blindness (MIB) effect and typical data showing its magnitude. Observers viewed high-contrast target patterns together with a dynamic random dot pattern ('rotating sphere') and reported the disappearance of the targets (see Methods). **a**, A typical snapshot of the dynamic display used. **b**, The percentage of accumulated invisibility period for the disappearance of one or more patches, exactly two patches and exactly three patches. The dots disappeared for about 30% of the trial duration, with disappearance episodes extending up to 10 s.

© 2001 Macmillan Magazines Ltd

nomenon of target dots that disappeared in one quadrant and reappeared in another after a few seconds. Such disappearance is unlikely to be explained by local adaptation or retinal stabilization effects. Third, the effect did not depend on local masking, as targets continued to disappear even when surrounded by background-colour circular 'protection zones' that occluded the moving mask (Fig. 2e). In such cases, targets disappeared without filling-in of the empty zones by the moving surround. However, targets did not disappear when positioned far outside the area of the mask (data not shown), suggesting that the effect is spatially limited. We also investigated the possibility that the effect depends on three-dimensional interpretation of the image; that is, on the structure from motion and occlusion. We found advantage for the three-dimensional dot sphere mask (Fig. 2h), although two-dimensional masks and Brownian motion could also induce the effect, leaving this issue open. Movement was critical in all cases but colour was not.

To investigate the functional level of visual processing affected by MIB, and to relate the MIB to other phenomena, we tested gestalt properties of figural organization (Figs 3 and 4). Good gestalts, defined by proximity or contour smoothness, tended to disappear entirely (as 'wholes') or to resist disappearance, as previously reported for stabilized images[16]. More surprising was the observed object rivalry that occurred between two partially overlapping (ellipses of different colours) or adjacent (orthogonal Gabor

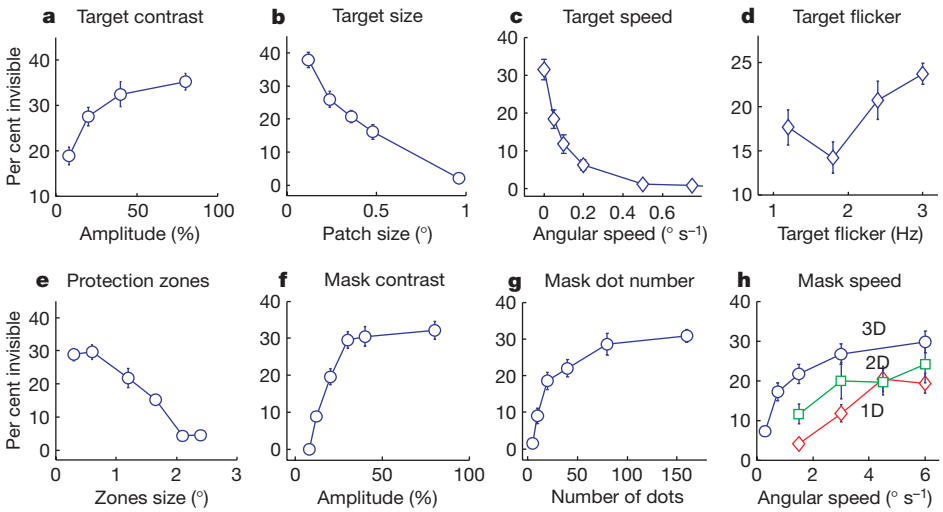


Figure 2 The effects of different parameters on the motion-induced blindness phenomenon. The percentage of accumulated invisibility period (any target) is plotted against different target and mask parameters (see Methods). **a**, Target luminance contrast. **b**, Target size. **c**, Target speed; the three-dot configuration was slowly rotated. **d**, Target flicker. **e**, The locality of the masking effect; size (diameter) of empty 'protection zones' around the targets. **f**, Mask luminance contrast. **g**, Mask number of dots. **h**, Mask speed and motion type: three-dimensional (3D) sphere (standard), two-dimensional (2D) rotation and one-dimensional (1D) left-to-right linear motion. Brighter targets disappeared more than dim ones (**a**), and disappearance was not eliminated with dynamic targets (**c**) or when the mask dots were distant from the target (**e**), and did not occur with static masks (**h**).

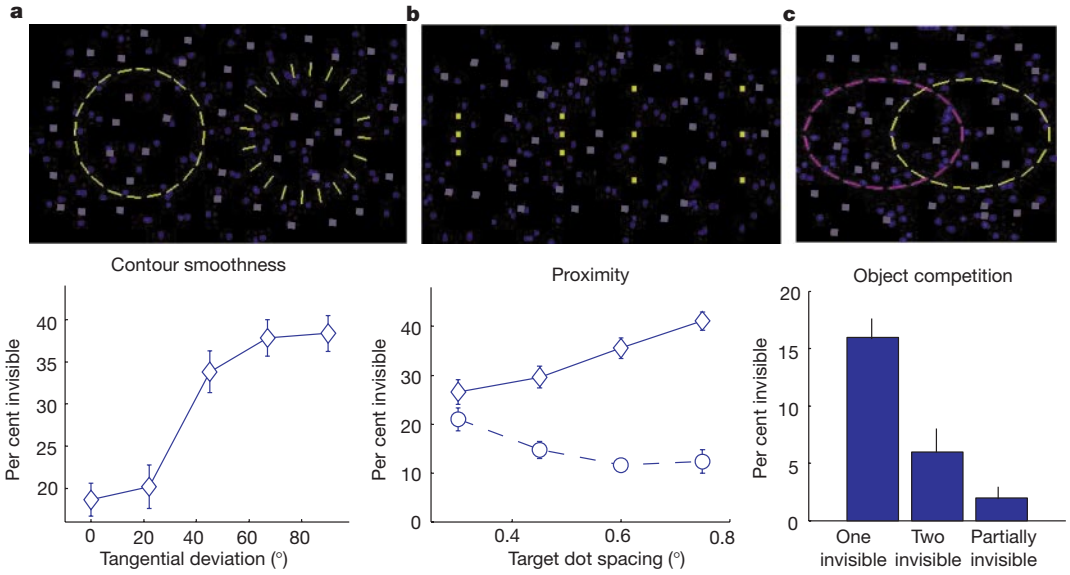


Figure 3 Gestalt effects reflected in the motion-induced blindness phenomena. The percentage of accumulated invisibility periods is plotted for different conditions. Snapshots of the two extreme conditions for each stimulus are presented at the top. The size of the mask dots was increased for clarity. **a**, Contour smoothness effect: the disappearance of any part of the circle is plotted against the tangential deviation of the circle elements, showing that smooth contours disappear less frequently. **b**, Proximity effect: the disappearance of any dot ('parts' condition) and of whole groups ('wholes' condition) are plotted against the interdot spacing. Increasing the spacing increased the disappearance of 'parts' but decreased the disappearance of 'wholes'. **c**, Object competition effect: complete and partial disappearance of each of two elliptical line configurations of different colours. Perception alternated between states of complete invisibility of each object for over 15% of the time.

© 2001 Macmillan Magazines Ltd

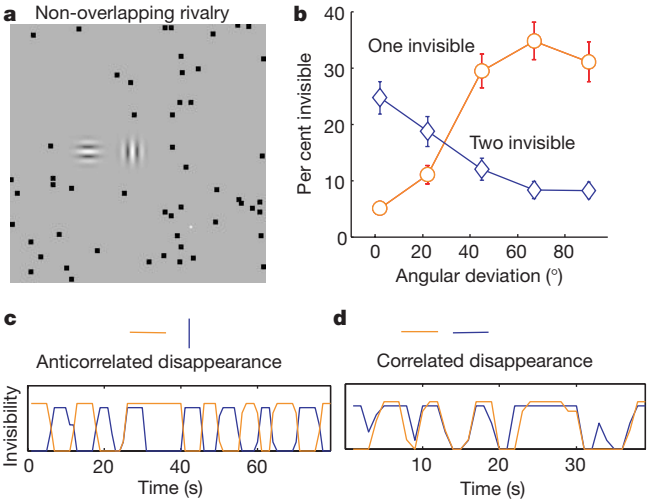


Figure 4 Co-operation and rivalry of non-overlapping stimuli under MIB conditions. **a**, A snapshot of stimuli used to measure anticorrelated disappearance (see Methods). **b**, Accumulated invisibility periods of one or two patches as a function of deviation from collinearity. With increased deviation, the percentage of 'one invisible' (rivalry) increased and the percentage of two invisible (co-operation) decreased. **c**, **d**, Invisibility against time (smoothed) as measured for one observer for orthogonal (**c**) and collinear (**d**) pairs of patches. The orthogonal patches engaged in anticorrelated disappearance (rivalry, $R = -0.75$), whereas the collinear patches were correlated (disappeared and reappeared together, $R = 0.76$).

patches) objects, which disappeared in alternation when superimposed by a 'rotating sphere' of dots. In Fig. 4c, we show an example of anticorrelated disappearance (rivalry) between two orthogonal patches compared with correlated disappearance for collinear patches (Fig. 4d). Overall, correlated (0.3–0.9) and anticorrelated or poorly correlated (−0.8 to 0.2) disappearance were observed for the collinear and orthogonal configurations respectively. Alternating disappearance of similar competing objects has been observed in simultanagnosia patients[10]. This suggests that MIB does not merely reflect suppression by movement but uncovers mechanisms of object competition that shift to operate in a winner-takes-all mode, as observed in the clinical cases.

We further observed that MIB is not uniformly distributed across space. Seven observers (two of them tested with each eye (see Methods)) reported more (typically twice as much) disappearance in the upper left field (two degrees of eccentricity) compared with the lower field. In comparison, we observed similar anisotropy in disappearance during figure–ground binocular rivalry, with targets presented to one eye and a static background mask (without motion) to the other (three observers). Moreover, two orthogonal Gabor patches presented to one eye, with a background of static random dot pattern presented to the other eye, often engaged in anticorrelated disappearance as in MIB. These similarities and the known disappearance of wholes and spatial configuration effects found in binocular[17–20] and monocular rivalry[8,18] suggest a common mechanism that gates or modulates conscious perception at the level of objects.

Why do salient stimuli disappear? Current explanations of other phenomena divide between sensory suppression[5,21] and 'higher-level' selection[22,23]. Our study shows that MIB is unlikely to be caused by sensory suppression or local adaptation. Recent evidence suggests the involvement of non-sensory or attention mechanisms. Using transcranial magnetic stimulation (TMS) with our basic stimuli (Fig. 1), it has been found[24] that a suppressive TMS pulse applied to parietal areas immediately after disappearance increased or decreased disappearance depending on the stimulated hemisphere. This suggests the involvement of attention mechanisms, which are assigned to competing objects and tend to divide between hemispheres. Additional evidence for the link between attention and disappearance comes from dorsal simultanagnosia patients, who report alternating disappearance of objects following bilateral occipito-parietal damage[10,11].

Taken together, these data suggest the following conclusions. (1) Under MIB conditions, or more generally for stimuli with sensory dissociation (different dynamics, eyes or colours), the visual system shifts to operate in a winner-takes-all mode. (2) This mode could be described as a disruption and slow-down of the commonly assumed, but usually unnoticed, fast attentional switching between objects in the scene. With such disruption, competing objects are perceived one at a time, with phenomenology similar to that observed in the clinical cases of attention deficits. (3) This disruption might occur because attentional mechanisms cannot be allocated or divided between dissociated or 'unfused' elements at the same time and location. (4) The actual rivalry and suppression could occur between competing object representations modulated by attention[25] or between attention mechanisms assigned to objects in space. The recent evidence for parietal mechanisms that represent visual space, and objects in space[26], as well as converging clinical evidence[3], might suggest the neural substrates for the mechanisms that gate visual appearance and disappearance. Finally, it is intriguing to consider the possibility that MIB, and visual disappearance in general, are just one manifestation of stimuli discarded by the visual system while fitting a consistent and useful interpretation to a fuzzy sensory input. In some rare cases, we are aware of the input being discarded, but most of the time we are not. □

Methods

Observers and stimuli

Ten observers (eight of them naive) took part in the experiments. In each experiment, four observers (three of them naive) viewed high-contrast patterns in a dark background together with a blue random dot (150 dots) pattern (mask) (see Supplementary Information). This pattern was displayed as if arranged on the surface of a 6° diameter rotating (x, y and z axes) sphere and viewed from 1 m for 60 s, repeated five times. In the experiments reported in Fig. 2, the targets consisted of three 0.2° yellow patches arranged along a 1° radius circle forming a triangle. Observers were instructed to attend to the rotating mask without following it with their eyes, while simultaneously reporting the disappearance of the targets by depressing three buttons, one for each target. Display luminance was set to 100 and 20 cd m^{-2} for 100% luminance contrast of the yellow and blue stimuli, respectively. Background luminance was set to 40 cd m^{-2} for the Gabor stimuli.

Parametric manipulations

Luminance contrast (Fig. 2a) varied in the range 10–80%. Size (Fig. 2b) varied in the diameter range 0.2–1°. Speed (Fig. 2c) was varied by slowly rotating the three-dot target configuration at angular speeds of 0–0.8° s^{-1}. Flicker (Fig. 2d) was varied by a linear modulation of target luminance between 0 and maximum luminance at 1–3 Hz. Locality of the masking effect was measured (Fig. 2e) by creating circular black 'protection zones'

© 2001 Macmillan Magazines Ltd

that occluded the mask dots with diameters of 0.2–2.5°. Mask luminance contrast (Fig. 2f) was varied in the range 10–80% and the number of mask dots (Fig. 2g) in the range 0–180. Mask speed and motion type were tested (Fig. 2h) by changing the angular speed (expressed as maximum visual angle displacement per s) for three types of motion: 3D sphere (standard); 2D rotation of an array of crosses (approximately the same size as the sphere); and 1D left-to-right linear motion of a similar array.

Gestalt effects

To test the contour smoothness effect (Fig. 3a), we varied the tangential deviation of line segments arranged in a 2.1° (diameter) circle, between smooth (0° deviation) and sun-shaped (90° deviation). Observers reported disappearance of any part of the circle.

To test the effect of proximity, we varied the spacing in two groups of three dots each, from 0.3° (Fig. 3b, left) to 0.75° (Fig. 3b, right). The groups were separated by 2.1° along the *x* axis. Observers reported the disappearance of any dot ('parts' condition) and the disappearance of whole groups ('wholes' condition).

To test object competition, we presented two elliptical line configurations (1.8°×2.2°, with 0.6° displacement) of different colours (Fig. 3c, top right), with observers reporting the complete disappearance of each configuration and partial disappearance of any part.

To test the co-operation and rivalry of non-overlapping stimuli under MIB conditions (Fig. 4), we used two adjacent Gabor patches. These had 8 cycles per degree and were three wavelengths apart, 2° from fixation in the upper left quadrant, on a grey background with a 'rotating sphere' made of black dots. They were shown to five observers.

To test the distribution of MIB across space, we presented a single yellow dot at 30 different locations within 5° from fixation, with a 5.5° radius sphere, three times for 45 s periods. These were shown to seven observers.

Received 12 February; accepted 17 April 2001.

1. Breese, B. B. On inhibition. *Psychol. Monogr.* **3,** 1–65 (1899).
2. Ditchburn, R. W. & Ginsborg, B. L. Vision with a stabilized retinal image. *Nature* **170,** 36–37 (1952).
3. Driver, J. & Vuilleumier, P. Perceptual awareness and its loss in unilateral neglect and extinction. *Cognition* **79,** 39–88 (2001).
4. O'Regan, J., Rensink, R. & Clark, J. Change-blindness as a result of 'mudsplashes'. *Nature* **398,** 34 (1999).
5. Blake, R. A neural theory of binocular rivalry. *Psychol. Rev.* **96,** 145–167 (1989).
6. Campbell, F. W., Glinsky, A. S., Howel, E. R., Riggs, L. A. & Atkinson, J. The dependence of monocular rivalry on orientation. *Perception* **2,** 123–125 (1973).
7. Andrews, T. J. & Purves, D. Similarities in normal and binocularly rivalrous viewing. *Proc. Natl Acad. Sci. USA* **94,** 9905–9908 (1997).
8. Bennet-Clark, H. & Evans, C. Fragmentation of patterned targets when viewed as prolonged after-images. *Nature* **199,** 1215–1216 (1963).
9. Troxler, D. in *Ophthalmologisches Bibliothek* (eds Himly, K. & Schmidt, J. A.) 51–53 (Fromman, Jena, 1804).
10. Luria, A. R. Disorders of 'simultaneous perception' in a case of bilateral occipito-parietal brain injury. *Brain* **82,** 437–449 (1959).
11. Rizzo, M. & Robin, D. A. Simultanagnosia: a defect of sustained attention yields insights on visual information processing. *Neurology* **40,** 447–455 (1990).
12. Grindley, G. C. & Townsend, V. Binocular masking induced by a moving object. *Q. J. Exp. Psychol.* **17,** 97–109 (1965).
13. Grindley, G. C. & Townsend, V. Further experiments on movement masking. *Q. J. Exp. Psychol.* **18,** 319–326 (1967).
14. Bonneh, Y., Cooperman, A. & Sagi, D. Loss of local pattern visibility in global shape perception. *Invest. Ophthalmol. Visual Sci.* (Suppl.) **40,** 4253 (1999).
15. Livingstone, M. S. & Hubel, D. H. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *J. Neurosci.* **7,** 3416–3468 (1987).
16. Evans, C. Some studies of pattern perception using a stabilized retinal image. *Br. J. Psychol.* **56,** 121–133 (1965).
17. Kovac's, I., Papathomas, T. V., Yang, M. & Feher, A. When the brain changes its mind: interocular grouping during binocular rivalry. *Proc. Natl Acad. Sci. USA* **93,** 15508–15511 (1996).
18. Walker, P. The perceptual fragmentation of unstabilized images. *Q. J. Exp. Psychol.* **28,** 35–45 (1976).
19. Fukuda, H. & Blake, R. Spatial interactions in binocular rivalry. *J. Exp. Psychol. Hum. Percept. Perform.* **18,** 362–370 (1992).
20. Bonneh, Y. & Sagi, D. Configuration saliency revealed in short duration binocular rivalry. *Vision Res.* **39,** 271–281 (1999).
21. Burbeck, C. & Kelly, D. Role of local adaptation in the fading of stabilized images. *J. Opt. Soc. Am.* **1,** 216–220 (1984).
22. Logothetis, N. K. Single units and conscious vision. *Phil. Trans. R. Soc. Lond. B* **353,** 1801–1818 (1998).
23. MacKay, D. M. in *Visual Neuroscience* (eds Pettigrew, J., Sanderson, K. & Levick, W.) 365–373 (Cambridge Univ. Press, Cambridge, New York, 1986).
24. Pettigrew, J. D. & Funk, A. P. Opposing effects on perceptual rivalry caused by right vs. left TMS. *Soc. Neurosci. Abstr.*(in the press).
25. Duncan, J., Humphreys, G. & Ward, R. Competitive brain activity in visual attention. *Curr. Opin. Neurobiol.* **7,** 255–261 (1997).
26. Colby, C. & Goldberg, M. Space and attention in parietal cortex. *Annu. Rev. Neurosci.* **22,** 319–349 (1999).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or the authors' web site (http://www.weizmann.ac.il/~masagi/MIB/mib.html).

Acknowledgements

We thank J. D. Mollon for recently drawing our attention to the work of Grindley and Townsend[13]. We thank E. Freeman, B. Zenger, S. Gepshtein, H. Reed, J. Pettigrew and M. Merzenich for their helpful comments.

Correspondence and requests for materials should be addressed to Y.S.B. (e-mail: yoram@ski.org).

Interaction with the NMDA receptor locks CaMKII in an active conformation

K.-Ulrich Bayer*, Paul De Koninck*†, A. Soren Leonard‡§, Johannes W. Hell‡§ & Howard Schulman*

* *Department of Neurobiology, Stanford University School of Medicine, Stanford, California 94305-5125, USA*

‡ *Department of Pharmacology, University of Wisconsin, Madison, Wisconsin 53706-1532, USA*

Calcium- and calmodulin-dependent protein kinase II (CaMKII) and glutamate receptors are integrally involved in forms of synaptic plasticity that may underlie learning and memory. In the simplest model for long-term potentiation[1], CaMKII is activated by Ca^{2+} influx through NMDA (*N*-methyl-D-aspartate) receptors and then potentiates synaptic efficacy by inducing synaptic insertion[2,3] and increased single-channel conductance[4] of AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptors. Here we show that regulated CaMKII interaction with two sites on the NMDA receptor subunit NR2B provides a mechanism for the glutamate-induced translocation of the kinase to the synapse in hippocampal neurons. This interaction can lead to additional forms of potentiation by: facilitated CaMKII response to synaptoc Ca^{2+}; suppression of inhibitory autophosphorylation of CaMKII; and, most notably, direct generation of sustained Ca^{2+}/calmodulin (CaM)-independent (autonomous) kinase activity by a mechanism that is independent of the phosphorylation state. Furthermore, the interaction leads to trapping of CaM that may reduce down-regulation of NMDA receptor activity[5]. CaMKII–NR2B interaction may be prototypical for direct activation of a kinase by its targeting protein.

The generation of autonomous CaMKII activity has an important role in various forms of synaptic plasticity, and has been regarded as 'molecular memory', as the kinase remains active after the initial Ca^{2+} stimulus has subsided (see refs 1, 6, 7 for review). Autonomous CaMKII activity depends on autophosphorylation of T286 in its auto-inhibitory domain. Autophosphorylation requires coincident binding of at least two Ca^{2+}/CaM molecules to a dodecameric CaMKII holoenzyme[8,9] and enables Ca^{2+}-spike-frequency decoding by the kinase[10]. T286 autophosphorylation also results in greatly enhanced CaM binding (CaM trapping)[11] by the highly abundant α-CaMKII, which may sequester CaM and limit its availability for the NMDA receptor and other synaptic proteins. The initial Ca^{2+} stimulus for CaMKII activation can be provided by the NMDA receptor, the only known activity-dependent binding partner for CaMKII at the synapse[12–14]. Binding of α-CaMKII to the NMDA receptor was reported to require autophosphorylation[14], whereas NMDA-stimulated translocation of the kinase to neuronal synapses does not[15,16]; thus the two events did not seem to be linked. Our binding studies, however, now demonstrate that stimulation by Ca^{2+}/CaM is sufficient to induce binding of α-CaMKII to the cytoplasmic carboxy terminus of NR2B (residues 839–1,482) and that autophosphorylation is not required. In fact, this NR2B domain contains two sites with different modes of regulated CaMKII binding (Fig. 1), a Ca^{2+}/CaM-regulated site within residues 1,120–1,482 of NR2B (NR2B-C) and a phosphorylation-regulated site within residues 839–1,120 (NR2B-P). CaMKII binds to NR2B-

† Present address: Centre de Recherche Université Laval Robert-Giffard, Beauport, Québec G1J 2G3, Canada.

§ Present address: Department of Physiology (A.S.L.) and Department of Pharmacology (J.W.H.), University of Iowa, Iowa City, Iowa 52246-1109, USA.

© 2001 Macmillan Magazines Ltd

P only if it is autophosphorylated at T286. Neither unphosphorylated wild-type kinase nor mock autophosphorylated T286A mutant bound this site even in the presence of Ca^{2+}/CaM. Autophosphorylated wild-type kinase binds to both sites even after removal of Ca^{2+}/CaM, and a T286D mutant mimics the binding characteristics of autophosphorylated kinase (Fig. 1). Most importantly, activation by Ca^{2+}/CaM alone, in the absence of ATP, is sufficient to induce interaction of wild type and T286A CaMKII with NR2B-C (Fig. 1).

We investigated whether the autophosphorylation-independent translocation of CaMKII observed in hippocampal neurons[15,16] could be reconstituted in a heterologous system by coexpression with NR2B. In unstimulated HEK cells, green fluorescent protein (GFP)-labelled CaMKII is evenly distributed throughout the cytoplasm. After generating a Ca^{2+} stimulus, both wild type and T286A mutant CaMKII translocate, but only when coexpressed with NR2B (Fig. 2a). However, a CaMKII mutant (I205K) that does not bind to NR2B *in vitro* (see Supplementary Information) does not show this movement even in the presence of NR2B (Fig. 2a). The I205K mutation does not reduce CaM affinity and allows normal kinase activation by Ca^{2+}/CaM[17]. Kinase and receptor are localized together after stimulation (Fig. 2b), further implicating a direct interaction. The autophosphorylation-deficient T286A mutant also co-localizes with the receptor, but the NR2B-binding-deficient I205K mutant does not (Fig. 2b). In hippocampal neurons, this I205K mutant also fails to translocate after stimulation with glutamate (Fig. 2c), a treatment that readily induced translocation of both wild type and T286A mutant kinase to postsynaptic sites (Fig. 2c; see also Supplementary Information). These data demonstrate that binding to NR2B can provide a mechanistic basis for the stimulation-induced translocation of CaMKII in neurons.

Although Ca^{2+}/CaM is necessary to initiate binding of CaMKII to NR2B-C, binding persists after washes with EGTA that effectively strip CaM from the kinase–receptor complex (Fig. 3a). During washes with Ca^{2+} but without CaM, however, CaM remained bound to the complex. Thus, interaction with NR2B-C seems to increase the CaM affinity of CaMKII, similar to CaM trapping by the autophosphorylated kinase, as CaM dissociates from unphosphorylated kinase that is not complexed with NR2B-C with an off rate of approximately 2 s^{-1} (ref. 11). To test this hypothesis and to narrow the region of NR2B-C involved, we examined the effects of various NR2B-derived peptides on the dissociation of an IAEDANS-labelled CaM (CaM^{I}) from CaMKII during a 'chase' with an excess of unlabelled CaM[11,18] (Fig. 3b). The short NR2B-C-derived peptide N2B-s (residues 1,289–1,310) reduced the dissociation rate of CaM^{I} from the kinase by about 20-fold. By contrast, the control peptide N2B-con (residues 1,095–1,119), which was derived from NR2B-P and has an amino-acid content similar to N2B-s, had little effect. The ability of the short N2B-s peptide to simulate the effect of the NR2B-C segment on CaM trapping suggests that it contains some or all of the residues within NR2B-C that anchor CaMKII to the NMDA receptor.

N2B-s is homologous to the segment surrounding T286 in the auto-inhibitory domain of the kinase (Fig. 3c) and may therefore mimic some of its interactions[17]. The T286 segment interacts with two sites on the kinase, and thus N2B-s could bind to either one of these sites (Fig. 3d). In the unstimulated or basal state, interaction of the T286 segment with a groove on the surface of the kinase (T site) serves to keep the kinase inactive by positioning an adjacent pseudo-substrate sequence in the substrate-binding site (S site)[17]. In the activated state of the kinase, Ca^{2+}/CaM displaces the T286 segment from the T site and enables it to interact with a vacant S site on a neighbouring subunit. NR2B-C (or N2B-s) might preferentially occupy either the vacated T or S site. Binding by means of the S site would result in inhibition of the kinase.

We tested the effect of CaMKII interaction with the NR2B-C fragment on kinase activity in a phosphorylation assay (Fig. 4a). NR2B-C-immobilized kinase retained about 70% of maximal Ca^{2+}/CaM-stimulated activity. Notably, the immobilized CaMKII remained active even after removal of Ca^{2+}/CaM with EGTA. This autonomous activity was 19% of the maximal Ca^{2+}/CaM-stimulated activity seen with NR2B-C-bound kinase. Such autonomous activity has previously been observed only after autophosphoryla-

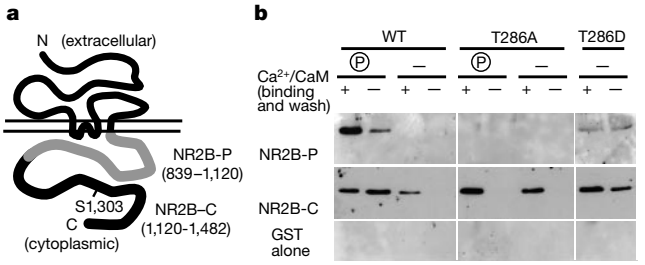


Figure 1 Differential regulation of CaMKII binding to two regions on NR2B. **a**, Schematic representation of the NR2B topology indicating the fragments used and the S1,301 phosphorylation site. **b**, Binding of α-CaMKII wild type (WT), T286A or T286D to immobilized fusion proteins of GST with NR2B-P or NR2B-C under various conditions. Kinase was either previously autophosphorylated or not (P or –), and the binding reactions and washes were done either in the presence or absence of Ca^{2+}/CaM as indicated. Bound CaMKII is visualized by immunoblotting the eluted protein using a specific antibody.

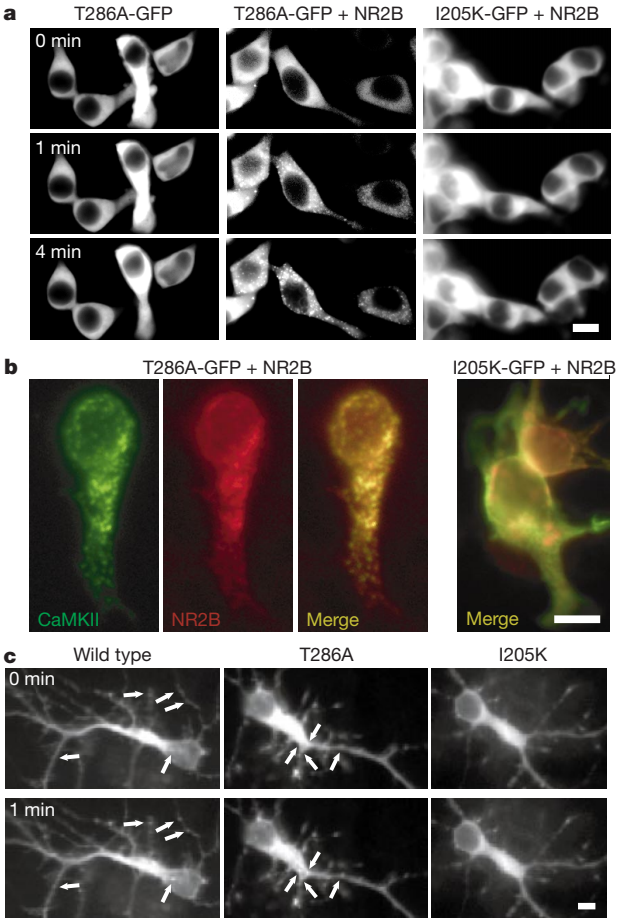


Figure 2 CaMKII translocation in HEK cells and in neurons. **a**, α-CaMKII-GFP T286A or I205K mutants were expressed in HEK cells, either alone or together with NR2B. The localization of the labelled kinase before and after a Ca^{2+} stimulus is monitored. **b**, Co-localization of α-CaMKII-GFP (green, GFP) and NR2B (red, immunostain) in stimulated HEK cells. **c**, Translocation of α-CaMKII-GFP in hippocampal neurons after stimulation with glutamate is monitored. Wild type (n = 8 of 12) and T286A mutant kinase (n = 2 of 4) rapidly translocates to postsynaptic sites (arrows; see Supplementary Information); the NR2B-binding deficient I205K mutant does not (n = 0 of 7). Scale bars, 10 μM.

© 2001 Macmillan Magazines Ltd

tion at T286, with maximal autonomy ranging between 20% and 80% (refs 7, 19). Immunoblotting with an antibody sensitive to phosphorylated T286 showed that the autonomously active NR2B-C-bound enzyme is not phosphorylated at T286 (data not shown, but see Fig. 5a). Indeed, even a CaMKII T286A mutant shows autonomous activity (33%) when bound to NR2B-C (Fig. 4a). Residual CaM cannot account for the autonomous activity as it was below the detectable level (see Fig. 3a), and is therefore at less than 1:50 stoichiometry to the kinase. The observed autonomous activity after NR2B-C binding is thus generated by a T286 phosphorylation-independent mechanism and is not due to trapping of CaM.

The most likely interpretation of the data above is that once NR2B-C occupies the T site of the kinase, it blocks interaction with the inhibitory T286 segment of that subunit so that the enzyme remains active even after dissociation of Ca^{2+}/CaM (Fig. 3d). As only some of the 12 subunits in each CaMKII holoenzyme[20] are likely to interact with NR2B, we tested the effect of soluble NR2B-derived peptides on CaMKII activity at concentrations that would saturate binding to all subunits (Fig. 4b, c). Peptides N2B-l (NR2B residues 1,259–1,310) and the shorter N2B-s are sufficient to generate autonomous activity of soluble CaMKII (Fig. 4b). The effect requires an initial exposure of the kinase to Ca^{2+}/CaM, as seen for NR2B-C binding. With increasing concentration of N2B-l, autonomous activity attains about 60% of initial maximal Ca^{2+}/CaM-stimulated activity (Fig. 4c), suggesting that N2B-l binds to most if not all kinase subunits with a direct activation of each one[8]. There is also some reduction in maximal stimulated activity at high concentration of peptide, which may result from secondary binding at the S site at such concentrations. This inhibitory effect is less prominent with the longer N2B-l peptide, which may have a greater preference for the T site owing to additional interactions. Phosphorylation of these peptides is not essential for this effect, as autonomy is also generated with peptide N2B-a in which S1,303 is replaced with alanine.

Can CaMKII that has been made partially autonomous by binding of some of its subunits to NR2B be further potentiated by T286 phosphorylation? Indeed, CaMKII bound to immobilized NR2B-C does phosphorylate T286, as detected by an antibody selective for phosphorylated T286, but only when stimulated by Ca^{2+}/CaM. Without Ca^{2+}/CaM, no T286 phosphorylation was observed even though phosphorylation of other residues does occur, as detected by ^{32}P incorporation (Fig. 5a). Phosphorylation of soluble CaMKII at T286 requires that CaM be bound to the subunit being phosphorylated to expose T286 for phosphorylation by a neighbouring subunit, a substrate-directed effect of CaM[8,9]. As

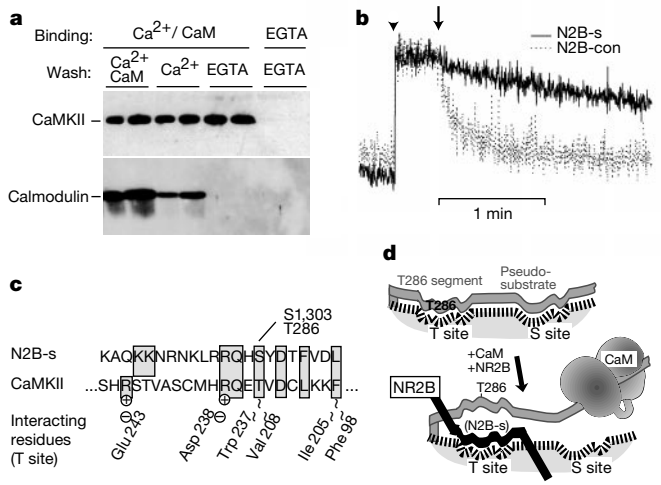


Figure 3 An NR2B fragment sufficient to induce CaM trapping by CaMKII. **a**, CaMKII incubated with immobilized NR2B-C in the presence of Ca^{2+}/CaM or EGTA, and washed with buffers containing EGTA, CaM and/or Ca^{2+}. Bound protein was eluted and tested for α-CaMKII and CaM by immunoblotting. **b**, CaM^{I} fluorescence increases on binding to CaMKII (ref. 18) (added with peptide present, arrowhead). Dissociation is initiated by the addition of excess unlabelled CaM (arrow), and is significantly decreased by 25 μM of N2B-s, but not by N2B-con. **c**, NR2B-derived CaMKII-binding peptide N2B-s and the auto-inhibitory region of CaMKII are homologous (boxed residues). CaMKII residues that interact with the auto-inhibitory region[17] are indicated. T286 of the kinase and S1303 of the receptor are marked. **d**, Model for CaMKII binding to NR2B. Binding of Ca^{2+}/CaM to individual CaMKII subunits displaces the auto-inhibitory region from the S and T sites (substrate binding and T286-segment-binding site) of the kinase. The vacated T site can then interact with the NR2B region around S1,303, preventing rebinding of the auto-inhibitory region and leaving the S site accessible for autonomous substrate phosphorylation.

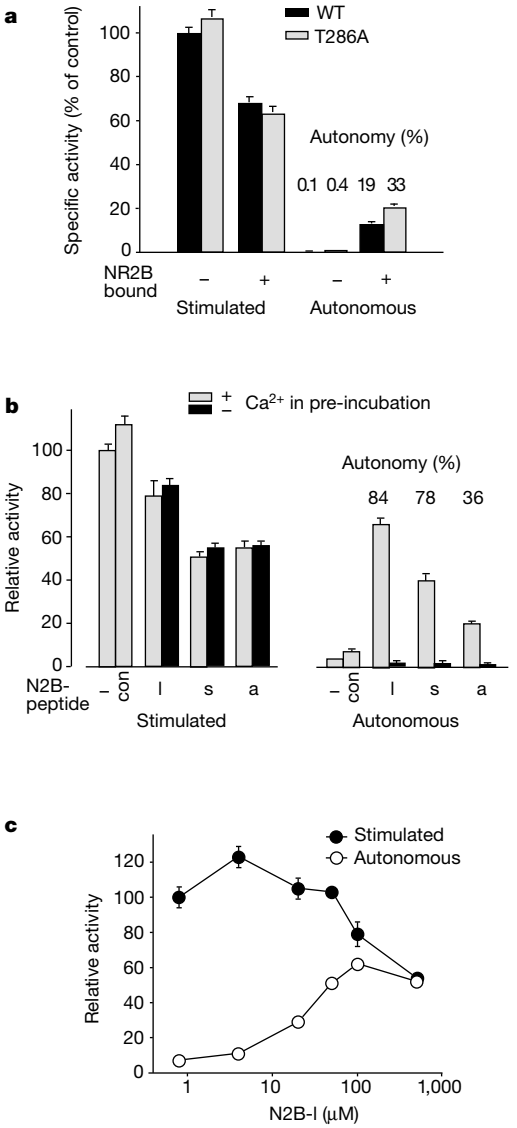


Figure 4 Interaction of CaMKII with NR2B-C generates autonomous kinase activity by a mechanism that does not require autophosphorylation or CaM trapping. **a**, α-CaMKII wild type (WT) or T286A mutant was bound to immobilized NR2B-C, and CaM was removed by EGTA washes (see Fig. 3a). Specific kinase activity of the NR2B-bound CaMKII was compared with soluble enzyme, either in the presence of Ca^{2+}/CaM (stimulated) or EGTA (autonomous). **b**, CaMKII was pre-incubated with NR2B-derived peptides (–, no peptide; or N2B-l, -s, -a or -con), CaM and either Ca^{2+} or EGTA (+ or – Ca^{2+}). EGTA was added to dissociate CaM from the kinase–peptide complexes, and after 8 min Ca^{2+}/CaM stimulated or autonomous activity was measured. **c**, Stimulated and autonomous kinase activity is dependent on the N2B-l concentration during pre-incubation (compare with **b**). For **a**–**c** The specific activity of Ca^{2+}/CaM-stimulated kinase (preincubation without peptide) was set as 100%. All data points are means ± s.e.m.

© 2001 Macmillan Magazines Ltd

NR2B-C should only bind one or a few of the subunits in each holoenzyme, it is likely that T286 is not exposed in the unbound subunits. We therefore activated every subunit by saturating the kinase with NR2B-derived peptides and tested for T286 autophosphorylation. Even full autonomy produced by the peptides did not lead to significant T286 phosphorylation in the absence of Ca^{2+}/CaM (Fig. 5b). However, once Ca^{2+}/CaM exposes T286, it is readily phosphorylated by peptide-bound subunits. Therefore, NR2B-C binding sustains kinase activity without CaM, but does not mimic the substrate-directed effect of CaM binding.

Binding of Ca^{2+}/CaM to CaMKII is inhibited by a 'burst' of autophosphorylation at T305 and T306, which is initiated when CaM dissociates from an autonomous enzyme[21,22]. We compared soluble CaMKII with NR2B-C-bound CaMKII under conditions known to produce the inhibitory burst (Fig. 5c). CaM binding to the soluble enzyme, as assessed by CaM overlay, is largely eliminated by the inhibitory burst of autophosphorylation. By contrast, the same conditions failed to reduce CaM binding for the NR2B-C-bound form of the kinase. Thus, binding to NR2B renders the kinase autonomous, as does T286 phosphorylation, but suppresses the inhibitory phosphorylation of T305/T306.

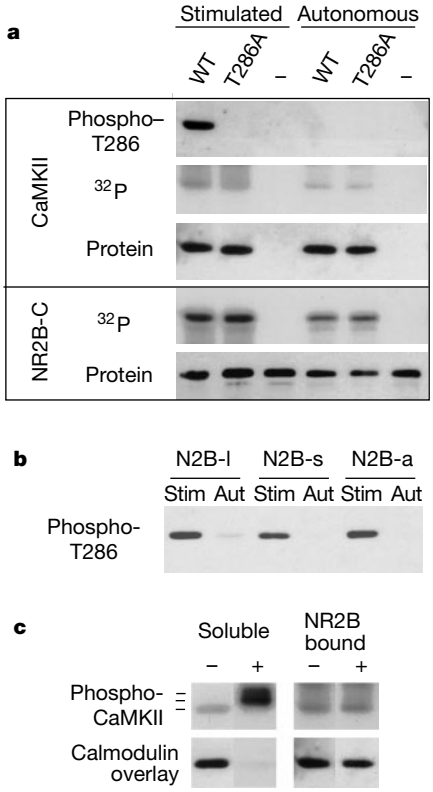


Figure 5 Phosphorylation within CaMKII–NR2B complexes. **a**, α-CaMKII wild type (WT), T286A or no kinase (–) bound to NR2B-C. Phosphorylation reactions contained either Ca^{2+}/CaM (stimulated) or EGTA (autonomous). Upper panels, phosphorylated CaMKII (^{32}P), T286-phosphorylated CaMKII (phospho-T286) and total CaMKII (protein), visualized by autoradiograph, immunoblotting with a phosphorylated T286-specific antibody, or CaM overlay. Lower panels, phosphorylated (^{32}P) and total NR2B-C (protein) visualized by autoradiography or immunoblotting. **b**, CaMKII was pre-incubated with peptides in the presence of Ca^{2+}/CaM. After addition of EGTA (8 min), samples were subjected to Ca^{2+}/CaM stimulated (stim) or autonomous (aut) auto-phosphorylation reactions. T286-phosphorylated CaMKII is visualized by immunoblotting with a phospho-specific antibody. **c**, Binding to NR2B can protect CaMKII from inhibitory burst autophosphorylation. CaMKII in solution and bound to NR2B-C are compared. As in **a**, T286 autophosphorylation was stimulated in the presence of [γ-^{32}P]ATP (–). EGTA was added and the reaction continued at 30 °C to induce burst phosphorylation (+). ^{32}P incorporation visualized by autoradiography (upper panel); burst phosphorylation leads to a bandshift of CaMKII. Burst phosphorylation at T305/306 inhibits CaM binding (CaM overlay, lower panel).

One possible substrate for CaMKII made autonomous by forming a kinase–receptor complex is the NMDA receptor itself[23]. CaMKII bound to NR2B-C phosphorylates the cytoplasmic domain of NR2B (Fig. 5a). In contrast to T286 autophosphorylation, NR2B-C phosphorylation occurs even without Ca^{2+}/CaM. However, Ca^{2+}/CaM further enhances the NR2B-C phosphorylation, probably owing to activation of non-autonomous unbound subunits. The phosphorylation of NR2B-C does not disrupt the interaction with the kinase, as no CaMKII was detected in the supernatant of the reactions (not shown). Ca^{2+} signals can stimulate both phosphorylation and dephosphorylation of NR2B[23,24], but the CaMKII–NR2B interaction generates a persistent kinase activity that outlasts the initial stimulus.

Stimulus-dependent binding to NR2B targets CaMKII to the signalling protein network of the postsynaptic density, where it is optimally positioned to decode local Ca^{2+} influx and mediate its proposed functions in learning and memory by modulating AMPA receptor-dependent neurotransmission[1–4], NMDA receptor inactivation[5,25], and the mitogen-activated protein kinase pathway[26]. In contrast to the previously identified CaMKII-anchoring proteins αKAP, F-actin and densin-180 (refs 27–30), NR2B affects not only the localization of the kinase but also directly regulates several of its physiologically important properties by mimicking a segment of its auto-inhibitory domain. In contrast to T286 phosphorylation, NR2B binding generates an autonomous and CaM-trapping state of CaMKII that cannot be reversed by phosphatases. It also suppresses inhibitory autophosphorylation at T305/306 that promote dissociation of the kinase from synaptic sites[16]. NR2B-induced CaMKII autonomy may preferentially occur with low-level stimuli as it does not require coincident binding of two CaM molecules[8,9], but may facilitate T286 autophosphorylation of neighbouring subunits on binding of a single CaM. In turn, autophosphorylation promotes CaMKII interaction with a second, phosphorylated T286-dependent binding site on NR2B. Thus, NR2B-binding and T286 autophosphorylation can function synergistically. Collectively, these mechanistic steps constitute a powerful feed-forward pathway for the regulation and potentiation of synaptic strength. □

Methods

Fluorescence-based CaM dissociation assay

Baculovirus-based purification of CaMKII and the CaM dissociation assay will be described in detail elsewhere (S. I. Singla, A. Hudmon, J. M. Goldberg, J. L. Smith and H. Schulman, submitted manuscript). The assays contained 15 nM CaM[I], 300 nM CaMKII, 50 mM MOPS pH 7, 0.1 mg ml^{-1} BSA, 150 mM KCl, 2 mM Mg-ADP, 0.5 mM $CaCl_2$ and (in the chase) 4 μM unlabelled CaM.

Binding assays with immobilized GST fusion protein

Glutathione *S*-transferase (GST)–NR2B fusion proteins were expressed in bacteria and immobilized on anti-GST-antibody-coated 96-well plates (Gibco, BRL). Binding mix (50 μl) (100–200 nM kinase, 50 mM PIPES pH 7.0, 0.1% BSA, 150 mM NaCl, 0.1% Tween-20, and either 1 mM $CaCl_2$ and 3 μM CaM or 1 mM EGTA) was overlaid for 1 h at 4 °C after blocking with 5% BSA. Wells were washed 7–8 times over a period of 20 min with buffer containing EGTA (0.5–1 mM) or as indicated (1 mM $CaCl_2$ ± 300 nM CaM). Immobilized kinase activity was measured for 1 min at 30 °C (50 ml of 50 mM PIPES pH 7, 0.1 mg ml^{-1} BSA, 10 mM $MgCl_2$, 100 μM [γ-^{32}P]ATP (1 Ci $mmol^{-1}$), 15 μM autocamtide-3 (AC3) peptide substrate, and either 0.5 mM $CaCl_2$ and 900 nM CaM or 0.5 mM EGTA) or the complexed proteins were phophorylated for 2 min on ice (similar buffers but without AC3). For immunoblot analysis using specific antibodies against α-CaMKII (Cbα2, Gibco), phosphorylated T286 (ABR) or CaM (Upstate Biotechnology), and for overlay with biotinylated CaM (STI), protein was eluted for 12 min in boiling SDS-loading buffer.

Translocation of CaMKII in HEK cells and neurons

Hippocampal cultures from neonatal rats were grown at high density (300–600 cells mm^{-2}) on 13 mm poly-D-lysine-coated aclar coverslips in neurobasal medium supplemented with B27, L-glutamine (500 μM), penicillin/streptomycin (50 U ml^{-1}/50 μg ml^{-1}) and cytosine arabinoside (5 μM, from day 2 to 5). Seven-day-old cultures were transfected in the presence of 10 μM CNQX (6-cyano-7-nitroquinoxaline-2,3-dione)/2-amino-5-phosphono pentanoate acid and 50 μM AP-5, using a custom-made microporator (platinum

© 2001 Macmillan Magazines Ltd

wires 7 mm apart) and a BioRad Gene Pulser (with Pulse Controller, 600 Ω; capacitance extender, 125 μF) set to deliver two 190 V pulses (30 s apart, at opposite polarity and with exponential decay τ-values of 30–40 ms). We injected 5 μl of expression vector (0.5 μg μl^{-1}) into the microporator before the pulses. Neurons were maintained for 16–24 h (in the presence of 25 μM AP-5) before imaging. HEK cells were transfected using lipofectamine Plus (BRL), passaged after 16–20 h onto coverslips and maintained for 6–24 h. Images were acquired on a Nikon inverted microscope with a ×40 (1.3 NA) oil immersion objective using a Pentamax cooled CCD camera controlled by Metamorph software (Universal Imaging). Cells were perfused with Hank's Balanced Salt Solution containing 1 mM $MgCl_2$ (for neurons), 1.2 mM $CaCl_2$ and 10–25 mM HEPES for 5–10 min at room temperature in a custom-made chamber. To induce a Ca^{2+} stimulus, HEK cells were treated with ionomycin/EGTA (10 μM/1 mM) for 2 min, and then perfused with 2 mM $CaCl_2$; neurons were stimulated with glutamate/glycine (100 μM/10 μM) (see ref. 15 and references therein). For immunostaining with NR2B antibody (Transduction Laboratories, 1:400; secondary antibody: GαM-Alexa594, Molecular Probes, 1:1000), cells were fixed after stimulation with 4% paraformaldehyde in 0.1 M phosphate buffer, and permeabilized with 0.1% Triton X-100 in PBS.

NR2B-derived peptides and CaMKII activity

The peptides N2B-l, -s, -a, and -con (NR2B 1,259–1,310, 1,289–1,310, 1,289–1,310/S1,303A and 1,095–1,119, respectively) were pre-incubated for 12 min with kinase (100 nM CaMKII, 100 μM peptide, 4.5 μM CaM, 10 mM PIPES pH 7.0, 0.1 mg ml^{-1} BSA, and 0.25 mM $CaCl_2$ or EGTA). EGTA was added (2.5 mM) and samples were diluted 1:5 into activity assays (see 'Binding assays', above, but with 100 μM AC3) or autophosphorylation reactions (without AC3).

Received 26 February; accepted 17 April 2001.

1. Malenka, R. C. & Nicoll, R. A. Long-term potentiation—a decade of progress? *Science* **285,** 1870–1874 (1999).
2. Rongo, C. & Kaplan, J. M. CaMKII regulates the density of central glutamatergic synapses *in vivo*. *Nature* **402,** 195–199 (1999).
3. Hayashi, Y. *et al.* Driving AMPA receptors into synapses by LTP and CaMKII: requirement for GluR1 and PDZ domain interaction. *Science* **287,** 2262–2267 (2000).
4. Derkach, V., Barria, A. & Soderling, T. R. Ca^{2+}/calmodulin-kinase II enhances channel conductance of alpha-amino-3-hydroxy-5-methyl-4-isoxazolepropionate type glutamate receptors. *Proc. Natl Acad. Sci. USA* **96,** 3269–3274 (1999).
5. Ehlers, M. D., Zhang, S., Bernhadt, J. P. & Huganir, R. L. Inactivation of NMDA receptors by direct interaction of calmodulin with the NR1 subunit. *Cell* **84,** 745–755 (1996).
6. Lisman, J., Malenka, R. C., Nicoll, R. A. & Malinow, R. Learning mechanisms: the case for CaM-KII. *Science* **276,** 2001–2002 (1997).
7. Schulman, H. & Brown, A. in *Calcium as Cellular Regulator* (eds. Carafoli, E. & Klee, C.) 311–343 (Oxford Univ. Press, New York, 1999).
8. Hanson, P. I., Meyer, T., Stryer, L. & Schulman, H. Dual role of calmodulin in autophosphorylation of multifunctional CaM kinase may underlie decoding of calcium signals. *Neuron* **12,** 943–956 (1994).
9. Rich, R. C. & Schulman, H. Substrate-directed function of calmodulin in autophosphorylation of Ca^{2+}/calmodulin-dependent protein kinase II. *J. Biol. Chem.* **273,** 28424–28429 (1998).
10. De Koninck, P. & Schulman, H. Sensitivity of CaM kinase II to the frequency of Ca^{2+} oscillations. *Science* **279,** 227–230 (1998).
11. Meyer, T., Hanson, P. I., Stryer, L. & Schulman, H. Calmodulin trapping by calcium-calmodulin-dependent protein kinase. *Science* **256,** 1199–1201 (1992).
12. Gardoni, F. *et al.* Calcium/calmodulin-dependent protein kinase II is associated with NR2A/B subunits of NMDA receptor in postsynaptic densities. *J. Neurochem.* **71,** 1733–1741 (1998).
13. Leonard, A. S., Lim, I. A., Hemsworth, D. E., Horne, M. C. & Hell, J. W. Calcium/calmodulin-dependent protein kinase II is associated with the *N*-methyl-D-aspartate receptor. *Proc. Natl Acad. Sci. USA* **96,** 3239–3244 (1999).
14. Strack, S., McNeill, R. B. & Colbran, R. J. Mechanism and regulation of calcium/calmodulin-dependent protein kinase II targeting to the NR2B subunit of the *N*-methyl-D-aspartate receptor. *J. Biol. Chem.* **275,** 23798–23806 (2000).
15. Shen, K. & Meyer, T. Dynamic control of CaMKII translocation and localization in hippocampal neurons by NMDA receptor stimulation. *Science* **284,** 162–166 (1999).
16. Shen, K., Teruel, M. N., Connor, J. H., Shenolikar, S. & Meyer, T. Molecular memory by reversible translocation of calcium/calmodulin-dependent protein kinase II. *Nature Neurosci.* **3,** 881–886 (2000).
17. Yang, E. & Schulman, H. Structural examination of autoregulation of multifunctional calcium/calmodulin-dependent protein kinase II. *J. .Biol. Chem.* **274,** 26199–26208 (1999).
18. Putkey, J. A. & Waxham, M. N. A peptide model for calmodulin trapping by calcium/calmodulin-dependent protein kinase II. *J. Biol. Chem.* **271,** 29619–29623 (1996).
19. Smith, M. K., Colbran, R. J., Brickey, D. A. & Soderling, T. R. Functional determinants in the autoinhibitory domain of calcium/calmodulin-dependent protein kinase II. Role of His282 and multiple basic residues. *J. Biol. Chem.* **267,** 1761–1768 (1992).
20. Kolodziej, S. J., Hudmon, A., Waxham, M. N. & Stoops, J. K. Three-dimensional reconstructions of calcium/calmodulin-dependent (CaM) kinase IIα and truncated CaM kinase IIα reveal a unique organization for its structural core and functional domains. *J Biol. Chem.* **275,** 14354–14359 (2000).
21. Colbran, R. J. & Soderling, T. R. Calcium/calmodulin-independent autophosphorylation sites of calcium/calmodulin-dependent protein kinase II. Studies on the effect of phosphorylation of threonine 305/306 and serine 314 on calmodulin binding using synthetic peptides. *J. Biol. Chem.* **265,** 11213–11219 (1990).
22. Hanson, P. I. & Schulman, H. Inhibitory autophosphorylation of multifunctional Ca^{2+}/calmodulin-dependent protein kinase analyzed by site-directed mutagenesis. *J. Biol. Chem.* **267,** 17216–17224 (1992).
23. Omkumar, R. V., Kiely, M. J., Rosenstein, A. J., Min, K. T. & Kennedy, M. B. Identification of a phosphorylation site for calcium/calmodulin-dependent protein kinase II in the NR2B subunit of the *N*-methyl-D-aspartate receptor. *J. Biol. Chem.* **271,** 31670–31678 (1996).
24. Lieberman, D. N. & Mody, I. Regulation of NMDA channel function by endogenous Ca^{2+}-dependent phosphatase. *Nature* **369,** 235–239 (1994).
25. Hisatsune, C. *et al.* Phosphorylation-dependent regulation of *N*-methyl-D-aspartate receptors by calmodulin. *J. Biol. Chem.* **272,** 20805–20810 (1997).
26. Chen, H. J., Rojas-Soto, M., Oguni, A. & Kennedy, M. B. A synaptic Ras-GTPase activating protein (p135 SynGAP) inhibited by CaM kinase II. *Neuron* **20,** 895–904 (1998).
27. Bayer, K.-U., Löhler, J. & Harbers, K. An alternative, nonkinase product of the brain-specifically expressed Ca^{2+}/calmodulin-dependent kinase II α isoform gene in skeletal muscle. *Mol. Cell. Biol.* **16,** 29–36 (1996).
28. Bayer, K.-U., Harbers, K. & Schulman, H. αKAP is an anchoring protein for a novel CaM kinase II isoform in skeletal muscle. *EMBO J.* **17,** 5598–5605 (1998).
29. Shen, K., Teruel, M. N., Subramanian, K. & Meyer, T. CaMKIIβ functions as an F-actin targeting module that localizes CaMKIIα/β heterooligomers to dendritic spines. *Neuron* **21,** 593–606 (1998).
30. Strack, S., Robison, A. J., Bass, M. A. & Colbran, R. J. Association of calcium/calmodulin-dependent kinase II with developmentally regulated splice variants of the postsynaptic density protein densin-180. *J. Biol. Chem.* **275,** 25061–25064 (2000).

Supplementary information is available on *Nature's* World-Wide web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank A. Hudmon, S. Singla, M. Meffert, P. Nef and T. Bartfai for helpful discussions, and B. Barres, R. C. Malenka, T. Meyer and D. Mochly-Rosen for comments on the manuscript. Purified CaMKII and IAEDANS-CaM were provided by A. Hudmon and S. Singla. We also thank T. Meyer and colleagues for providing the α-CaMKII-GFP plasmid and help in designing the microporation technique. The research was supported by NIH grants (to H.S. and J.W.H.), a grant from Hoffmann-La Roche (H.S.), and an American Heart Award (J.W.H.). P.D.K. is a Career Awardee of the Burroughs Wellcome Fund and a Centennial Fellow of the Canadian Institutes of Health Research.

Correspondence and requests for materials should be addressed to K.-U.B. (e-mail: ulli.bayer@stanford.edu) or H.S. (e-mail: schulman@cmgm.stanford.edu).

Molecular mechanism of cAMP modulation of HCN pacemaker channels

Brian J. Wainger*, Matthew DeGennaro†, Bina Santoro*, Steven A. Siegelbaum*†‡ & Gareth R. Tibbs‡§

** Center for Neurobiology and Behavior; † Howard Hughes Medical Institute; ‡ Department of Pharmacology; and § Department of Anesthesiology, Columbia University, New York, NY 10032, USA*

Hyperpolarization-activated cation channels of the HCN gene family[1–6] contribute to spontaneous rhythmic activity in both heart[7] and brain[5,6,8]. All four family members contain both a core transmembrane segment domain, homologous to the S1–S6 regions of voltage-gated K^+ channels, and a carboxy-terminal 120 amino-acid cyclic nucleotide-binding domain (CNBD) motif. Homologous CNBDs are responsible for the direct activation of cyclic nucleotide-gated channels and for modulation of the HERG voltage-gated K^+ channel—important for visual and olfactory signalling[9] and for cardiac repolarization[10], respectively. The direct binding of cyclic AMP to the cytoplasmic site on HCN channels permits the channels to open more rapidly and completely after repolarization of the action potential[1,2,11], thereby accelerating rhythmogenesis[6–8]. However, the mechanism by which cAMP binding modulates HCN channel gating and the basis for functional differences between HCN isoforms remain unknown. Here we demonstrate by constructing truncation mutants that the CNBD inhibits activation of the core transmembrane domain. cAMP binding relieves this inhibition. Differences in activation gating and extent of cAMP modulation between the HCN1 and HCN2 isoforms result largely from differences in the efficacy of CNBD inhibition.

© 2001 Macmillan Magazines Ltd

HCN1 and HCN2 channels generate distinct hyperpolarization-activated currents that differ in their kinetics, steady-state voltage-dependence of activation, and extent of modulation by cAMP[1,2,12]. HCN1 channels activate more rapidly on hyperpolarization than HCN2 channels; additionally, HCN1 channels turn on at voltages of about 20 mV more positive than HCN2 channels[12] (Fig. 1a–c). Fits of tail-current-activation curves by the Boltzmann equation (see Methods) yielded a midpoint of activation ($V_{1/2}$) for HCN1 channels of –110.6 ± 1.3 mV (mean ± s.e.m.; $n = 9$) with a slope of 5.3 ± 0.2 mV (corresponding to an effective valence of 4.7 electronic

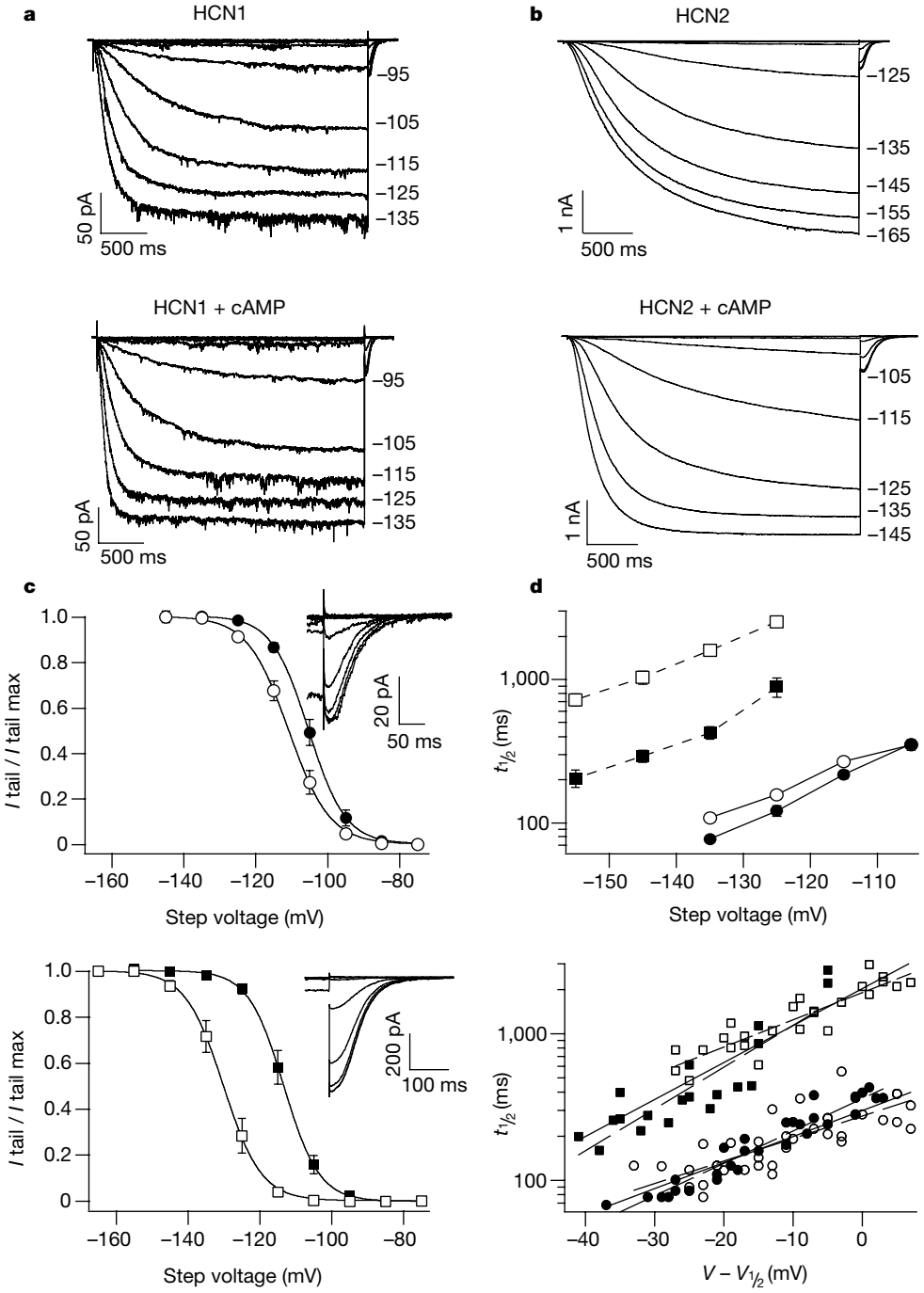


Figure 1 HCN1 and HCN2 channels differ in steady-state voltage dependence, kinetics of activation, and response to cAMP. **a**, **b**, Currents from inside-out patches containing either HCN1 (**a**) or HCN2 (**b**) channels. Top, currents in absence of cAMP; bottom, currents with cAMP (30 and 3 μM in **a** and **b**, respectively) in the solution bathing the internal side of the membrane. Currents were elicited by 3-s hyperpolarizations from a holding potential of –30 mV to a range of test potentials. **c**, Mean tail current *I*/*V* curves. Normalized peak inward tail-current amplitude as a function of voltage during preceding hyperpolarization for HCN1 (top panel, circles) and HCN2 (bottom panel, squares) channels in absence (open symbols) or presence (filled symbols) of cAMP. Insets, expanded records of HCN1 and HCN2 tail currents measured at –40 mV from **a** and **b** (top panels for both). Lines show fit of Boltzmann curve. Mean of 9 patches for HCN1 and 7 patches for HCN2. **d**, Relation between activation kinetics and voltage. $t_{1/2}$ values plotted as function of hyperpolarizing voltage for HCN1 (solid lines, circles) and HCN2 (dashed lines, squares) in the absence (open symbols) and presence (filled symbols) of cAMP (top panel). Relationship between activation kinetics and voltage relative to $V_{1/2}$ (bottom panel). $t_{1/2}$ values plotted as a function of difference between hyperpolarizing step voltage and the $V_{1/2}$ ($V - V_{1/2}$) for a particular experiment. Dashed lines, linear regressions for the four populations: HCN1 (open circles), HCN1 + cAMP (filled circles), HCN2 (open squares) and HCN2 + cAMP (filled squares). Solid lines, linear regressions for HCN1 (lower) and HCN2 (upper), combining data in the absence and presence of cAMP.

© 2001 Macmillan Magazines Ltd

charges). In contrast, HCN2 channels displayed a more negative midpoint of activation (−129.6 ± 1.9 mV (n = 7)) and a slope of 4.5 mV ± 0.3 mV (5.6 charges). For both HCN1 and HCN2 channels, the time-course of activation was a steep function of test potential (Fig. 1a, b, d), with the half time of activation ($t_{1/2}$) decreasing e-fold for every 20–30 mV hyperpolarization. At any given voltage, HCN2 activation was about 15-fold slower than HCN1 activation (Fig. 1d, top panel).

Application of cAMP to the cytoplasmic side of the membrane facilitated the opening of HCN2 to a much greater extent than HCN1 (Fig. 1a, b). cAMP shifted the $V_{1/2}$ of HCN2 by +17.3 ± 1.0 mV compared to a +5.8 ± 0.6 mV shift of HCN1. Moreover, cAMP accelerated HCN2 activation (3.5-fold) much more than HCN1 activation (1.5-fold; Fig. 1d, top panel). For both HCN1 and HCN2, the increase in activation rate with cAMP can be attributed to the shift in voltage-dependence of activation. Thus, $t_{1/2}$ values for a given channel isoform in the presence and absence of cAMP overlap when plotted as a function of voltage relative to the $V_{1/2}$ for each individual experiment (Fig. 1d, bottom panel). The larger increase in the rate of activation of HCN2 versus HCN1 with cAMP therefore reflects the more pronounced shift in $V_{1/2}$ of HCN2 compared to HCN1. Even when differences in $V_{1/2}$ are taken into account in this manner, however, HCN1 channels still activate 4–6-fold faster than HCN2 channels (as seen by the vertical displacement of the HCN1 and HCN2 plots in Fig. 1d, bottom).

We considered the following questions: what is the molecular mechanism by which cAMP modulates the gating of HCN1 and HCN2 channels? What accounts for differences in HCN1 and HCN2 gating in the absence of cAMP, and why does cAMP modulate HCN2 more robustly than HCN1? Is the hyperpolarization-activation phenotype of HCN channels determined uniquely by the core transmembrane domain of the channel, or does this behaviour require the influence of the CNBD? To investigate these questions, we constructed a series of HCN1 and HCN2 C-terminal truncation mutants. Deletion of the extreme C-terminal portion of the channel (residues distal to the CNBD) had little effect on gating or response to cAMP in both the HCN1 and HCN2 backgrounds ($HCN1_{\Delta C\text{-}X}$, $HCN2_{\Delta C\text{-}X}$) (Fig. 2). We next focused on larger truncations that extended into the CNBD.

On the basis of the X-ray crystal structure of the homologous CNBD from the bacterial catabolite-activating protein (CAP)[13], the HCN channel CNBD has a bipartite structure: a β-roll subdomain (consisting of an eight-stranded β-roll flanked by two short α-helixes) and a long C helix[5,6]. Experiments on cyclic nucleotide-gated (CNG) channels show that the C helix interacts with the purine ring of the ligand preferentially in the open state of the channel, contributing coupling energy that enhances channel opening[14,15]. In contrast, interaction of the β-roll with the cyclized phosphate of the ligand determines binding affinity but not coupling energy[16].

Consistent with the importance of the C helix in CNG channels, truncations of HCN2 that eliminated the C helix ($HCN2_{\Delta\text{-}C\alpha}$) eliminated the response to cAMP (Fig. 2). This deletion had little effect on voltage gating in the absence of cyclic nucleotide, as seen by the unchanged value of $V_{1/2}$ (Fig. 2). In contrast, deletion of the entire CNBD ($HCN2_{\Delta CNBD}$) not only eliminated the response to cAMP but also shifted the midpoint of activation by 24 mV to more depolarized potentials (Fig. 2). Notably, the $V_{1/2}$ value of $HCN2_{\Delta CNBD}$ was similar to that of wild-type HCN2 channels in the presence of cAMP.

The results from these deletions suggest a simple model for HCN modulation by cAMP in which the β-roll subdomain of the CNBD exerts an inhibitory effect on channel activation, shifting gating to more negative potentials (see Fig. 4 below). Binding of cAMP—or deletion of the entire CNBD—would relieve this inhibition and shift gating to more positive potentials. Thus, $HCN2_{\Delta\text{-}C\alpha}$ remains trapped in the inhibited state because it contains the inhibitory proximal region of the CNBD but cannot respond to ligand owing to removal of the C helix.

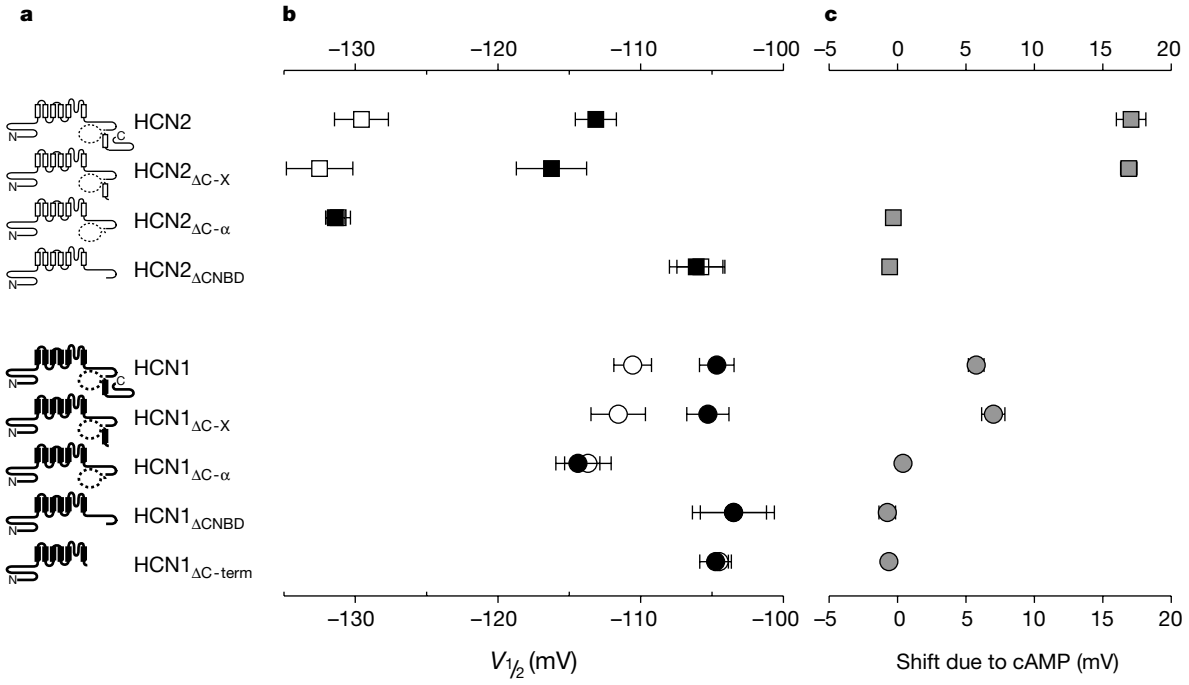


Figure 2 Effect of C-terminal truncation on steady-state voltage dependence in the presence and absence of cAMP. **a**, Schematic of different constructs. Thin lines, HCN2 constructs; thick lines, HCN1 constructs. The core transmembrane domain consists of the S1–S6 helixes (rectangles) and the P-region, a loop connecting the S5 and S6 helixes. Broken lines indicate the A and B helices and β-roll of the CNBD. The rectangle C-terminal to these structures represents the C helix of the CNBD. **b**, $V_{1/2}$ of deletion mutants in the absence (open symbols) and presence (filled symbols) of cAMP. Numbers of patches were (from top to bottom): 7, 4, 5, 9, 9, 7, 7, 4 and 4. **c**, Shift in $V_{1/2}$ value in response to cAMP. For each cAMP application, $V_{1/2}$ in the presence of cAMP was subtracted from the average of $V_{1/2}$ measurements before cAMP application and after washout of cAMP.

© 2001 Macmillan Magazines Ltd

Given the smaller response to cAMP of HCN1 channels compared with HCN2 channels, our model predicts that deletion of the CNBD of HCN1 would have a qualitatively similar but quantitatively smaller effect on gating than seen in the HCN2 background. Consistent with these predictions, we found that deletion of the entire CNBD of HCN1 ($HCN1_{\Delta CNBD}$) both eliminated the response to cAMP and shifted the $V_{1/2}$ of the steady-state activation curve by 5–7 mV. This shift was much smaller than the shift seen on truncation of the HCN2 CNBD, but similar to the 5 mV shift in $V_{1/2}$ for HCN1 channels in response to cAMP (Fig. 2). As observed for HCN2, deletion of the C helix of HCN1 ($HCN1_{\Delta\text{-}C\alpha}$) eliminated the response to cAMP while causing little change in steady-state activation gating (Fig. 2). The steady-state activation curves for $HCN1_{\Delta CNBD}$ and $HCN2_{\Delta CNBD}$ were nearly identical ($V_{1/2}$ values of −103.5 versus −105.8 mV, respectively), whereas the $V_{1/2}$ values of wild-type HCN1 and HCN2 channels differed by approximately 20 mV. These results suggest that differences in basal voltage dependence, as well as cAMP modulation, between HCN1 and HCN2 are due to differential inhibition by the CNBDs: in HCN2 channels the inhibition is greater than in HCN1. Consequently, deletion of the binding domain or application of cAMP produces a larger voltage shift in HCN2 than in HCN1.

A similar inhibitory mechanism is likely to occur in native channels. In cardiac myocytes, internal pronase shifted activation gating of hyperpolarization-activated currents by +57 mV and eliminated the response to cAMP[17] (although the very large voltage shift with pronase suggests some additional action distinct from the C terminus).

In both HCN and CNG channels, a highly conserved C-linker region of about 80 amino acids connects the last transmembrane segment (S6) to the CNBD. In CNG channels, this region couples conformational changes in the ligand-binding domain to channel activation[18–20]. We examined the contribution of the C linker to HCN channel function by constructing a deletion mutant that lacked all but three of the 522 amino acids of its C terminus ($HCN_{\Delta C\text{-}term}$). Notably, the $HCN1_{\Delta C\text{-}term}$ mutant formed functional channels that generated robust hyperpolarization-activated currents (the $HCN2_{\Delta C\text{-}term}$ mutant did not show any functional expression). The gating of $HCN1_{\Delta C\text{-}term}$, which lacks a C linker, was identical to that of $HCN1_{\Delta CNBD}$, which contains a C linker (Fig. 2). Thus, neither the C linker nor any other domain of the C terminus is necessary for hyperpolarization gating. Moreover, truncations that deleted the first half of the cytoplasmic amino terminus also had little effect on gating in HCN2 channels[21] and little effect on either gating or cAMP modulation in HCN1 channels (B. J. Wainger *et al.*, unpublished observations) further delineating the central function of the core transmembrane domain.

If cAMP binding acts to relieve CNBD-dependent inhibition, deletion of the CNBD should also speed activation kinetics, in a manner similar to cAMP. $HCN2_{\Delta CNBD}$ did indeed show a marked (fourfold) increase in activation rate relative to wild-type HCN2, resembling (but slightly greater than) the effect of cAMP on wild-

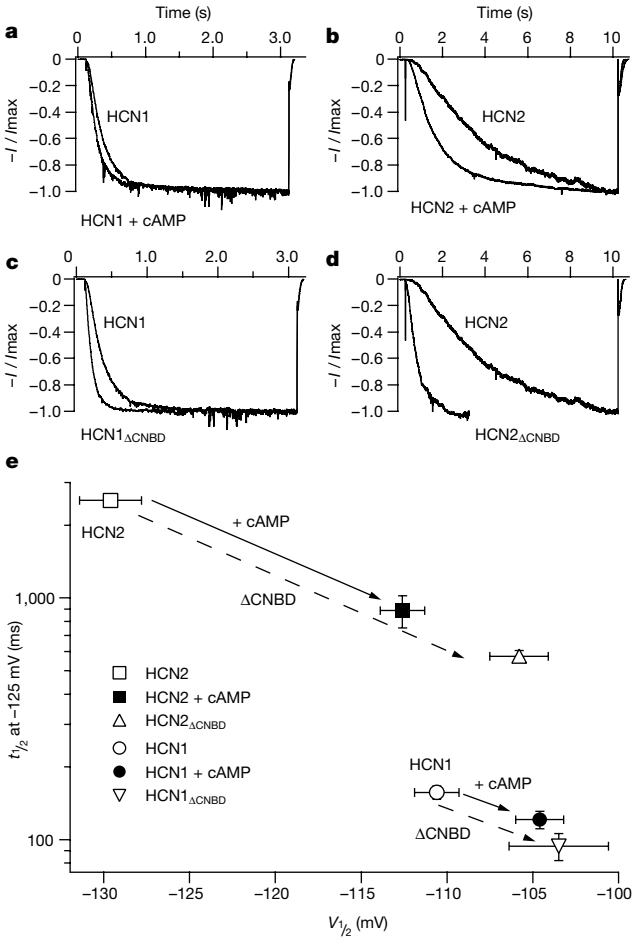


Figure 3 Deletion of CNBD mimics cAMP-induced increase in activation rate. **a**, **b**, Superimposed current records (normalized to maximal current amplitude during the pulse, I_{max}) showing effect of cAMP on HCN1 (**a**) and HCN2 (**b**) currents elicited by either 3-s (**a**) or 10-s (**b**) voltage steps to −125 mV. **c**, **d**, Comparison of currents for full-length HCN channels and HCN CNBD deletions during steps to −125 mV in both HCN1 (**c**) and HCN2 (**d**) backgrounds (in absence of cAMP). Voltage steps were 3 s for both deletion constructs; voltage step for HCN2 was 10 s. **e**, Comparison of effects on $V_{1/2}$ and $t_{1/2}$ values on application of cAMP (solid line) or deletion of the CNBD (dashed line) for HCN2 and HCN1.

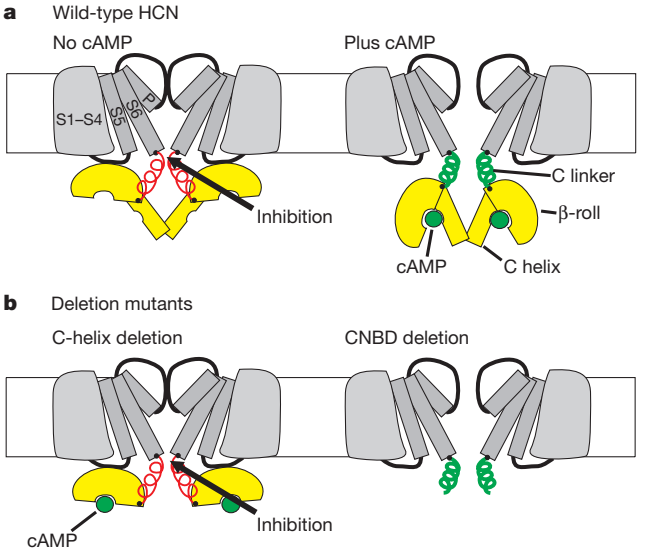


Figure 4 Model for cAMP modulation of HCN channel gating. **a**, Effect of cAMP on the wild-type HCN channel. The diagram illustrates two (adjacent) of the presumed four subunits that comprise an HCN homomeric channel. The core transmembrane domain (S1–S6) is indicated in grey. The coils represent the cytoplasmic C linker that couples the S6 transmembrane segment to the cyclic nucleotide-binding domain (CNBD), consisting of a β-roll subdomain (yellow semicircular region) and C helix (yellow cylinder). Left, in the absence of cAMP, the β-roll subdomain inhibits the gating of the core transmembrane domain by a strain coupled through the C linker (indicated in red). Right, in the presence of cAMP (green circle), a conformational change in the CNBD leads to relief of strain on the C linker (indicated in green), which promotes channel opening. **b**, Left, deletion of the C helix prevents cAMP binding from causing a conformational change in the CNBD to relieve inhibition. Right, deletion of the entire CNBD relieves strain on the C linker and relieves inhibition of gating.

© 2001 Macmillan Magazines Ltd

type HCN2 channels (Fig. 3b, d, e). $HCN1_{\Delta CNBD}$ showed a smaller (1.5-fold) increase in rate of activation relative to HCN1, mirroring the small effect of cAMP on full-length HCN1 (Fig. 3a, c, e). Thus, deletion of the CNBD resulted in a shift in both equilibrium ($V_{1/2}$) and kinetic ($t_{1/2}$) parameters that paralleled—although slightly exceeded—the effect of cAMP. The larger effect of the deletions may reflect the incomplete efficacy of agonist in relieving the inhibitory effect of the CNBD. Despite their nearly identical $V_{1/2}$ values, $HCN1_{\Delta CNBD}$ activated 5–6-fold faster than $HCN2_{\Delta CNBD}$ (Fig. 3e). This shows that the opening of the core HCN2 channel is inherently slower than HCN1, consistent with our finding that there is a 4–6-fold difference in kinetics between wild-type HCN1 and HCN2 after correcting for differences in their $V_{1/2}$ values (see Fig. 1d, bottom).

Our study demonstrates that the C-terminal CNBD inhibits HCN channel gating and that cAMP binding relieves this inhibition. On the basis of results in CNG channels[18–20] and preliminary studies on chimaeric HCN channels[22], the inhibitory action of the CNBD complex is probably coupled to the core domain through the C linker (Fig. 4). The C linker may constrain movement of the S6 helix, which is thought to form the activation gate in HCN[23] and voltage-gated K^+ channels[24]. Binding of cAMP would cause a conformational change in the CNBD that relieves this strain, allowing S6 gating to proceed unimpeded. Submembrane modulatory complexes that interact with the core domain of a channel have been demonstrated with the crystal structures of the N-terminal T1 assembly domain and the β-subunit of voltage-gated K^+ channels[25–27]. Given the ability of the core domain of the HCN channel to function on its own, cyclic nucleotide gating and modulation seem to take advantage of a general, evolutionarily conserved mechanism for coupling modulatory domains or subunits onto a central transmembrane core region that is responsible for channel gating. □

Methods

Truncation mutant construction and expression

Murine HCN1 and HCN2 were subcloned into pGH19 and pGHE expression vectors[28], respectively, using 5′ *Eco*RI and 3′ *Xba*I restriction enzyme sites. We made truncation constructs by introducing premature stop codons and *Xba*I sites by PCR, except for $HCN1_{\Delta C\text{-term}}$, for which we used an oligonucleotide linker containing a stop codon and an *Xba*I site. Fragments containing the mutations were subcloned into the parent channel using an endogenous *Pfl*MI site and the *Xba*I site. We sequenced regions generated by PCR or linker at least twice. All constructs were transcribed from *Nhe*I-linearized (HCN1 constructs) or *Sph*I-linearized (HCN2 constructs) DNA using T7 RNA polymerase (Message Machine; Ambion, Texas). We injected 50 ng RNA into each *Xenopus* oocyte.

We constructed the following mutants. $HCN1_{\Delta C\text{-}X}$: deletion of the extreme C terminus of HCN1. A stop codon was introduced after residue I610V (L611stop). $HCN1_{\Delta\text{-}C\alpha}$: deletion of the extreme C terminus and 25 amino-acid C helix of HCN1 (Y565stop). $HCN1_{\Delta CNBD}$: deletion of the CNBD and extreme C terminus of HCN1 (V473stop). $HCN1_{\Delta C\text{-term}}$: deletion of the entire C terminus (C linker, CNBD and extreme C terminus) of HCN1 (S391stop). $HCN2_{\Delta C\text{-}X}$: deletion of the extreme C terminus of HCN2 (Q665stop). $HCN2_{\Delta\text{-}C\alpha}$: deletion of the extreme C terminus and 25 amino-acid C helix of HCN2 (Y618stop). $HCN2_{\Delta CNBD}$: deletion of the CNBD and extreme C terminus of HCN2 (V526stop).

Electrophysiological recordings

Currents from cell-free inside-out patches were recorded 2–5 days after RNA injection[1]. Data were digitized and acquired using an ITC-16 interface (Instrutech) and acquired and analysed with Pulse and PulseFit software (Heka Electronics). Data were filtered at 1 kHz and sampled at 2 kHz. The pipette solution contained (in mM): 107 KCl, 5 NaCl, 10 HEPES and 2 $MgCl_2$, pH 7.4 with KOH; the bath solution contained 107 KCl, 5 NaCl, 10 HEPES, 1 $MgCl_2$ and 1 EGTA, pH 7.4 with KOH. The Ag–AgCl ground wire was connected to the bath solution by a 3 M KCl agar bridge electrode. We obtained all recordings at room temperature (22–24 °C). Because the midpoint of activation of both HCN1 and HCN2 shifted by approximately 40 mV in a hyperpolarized direction after patch excision, we waited at least 10 min after obtaining inside-out patches before beginning experiments. In most cases, this was sufficient to limit drift to a few millivolts over the time course of the experiment. We performed analysis of $V_{1/2}$ on patches that lasted through cAMP application. When indicated, we applied saturating concentrations of cAMP (mostly 3 μM, although 30 μM was occasionally used) on the intracellular side of patches.

Analysis

Tail current amplitudes were measured at −40 mV after the decay of capacitive transients. Tail current I/V curves were fitted by the Boltzmann equation,

$$I(V) = \mathrm{A1} + \mathrm{A2}/\{1 + \exp[(V - V_{1/2})/s]\}$$

in which A1 is the offset owing to holding current, A2 is the maximal tail current amplitude, V is the test pulse voltage, $V_{1/2}$ is the midpoint voltage of activation, and s is the slope. Mean tail current I/V curves were obtained by first subtracting the offset (A1) and then normalizing by the maximal amplitude (A2) I/V data from individual experiments. These normalized data were then averaged among different experiments and the group data fitted with a new Boltzmann curve. All shown $V_{1/2}$ data were determined from tail currents after 3-s hyperpolarizations (extending pulse duration to 10 s did not alter the $V_{1/2}$ value for the slowest construct). Standard error bars are shown in Figs 1–3.

Activation kinetics were determined from the half time of activation ($t_{1/2}$). Either 10-s or 3-s hyperpolarizing steps were used depending on the kinetics of the construct to permit activation to approach a steady state by the end of the pulse. For HCN2, which displayed the slowest kinetics, the 10-s pulses were sufficient for a steady state to be reached in only five out of eight experiments at −125 mV. The results from the other three experiments, in which a steady state was not reached, were not included in the analysis for Fig. 3. As a consequence, the reported $t_{1/2}$ value for HCN2 represents a lower limit for the true value.

With some measurements at −135, −145 and −155 mV of HCN2 activation in the presence of cAMP, 3-s pulses allowed the currents to approach steady state and were therefore included in our analysis. Regression analysis in Fig. 1d (top panel) was performed in SigmaPlot5.

Received 28 February; accepted 27 March 2001.

1. Santoro, B. *et al.* Identification of a gene encoding a hyperpolarization-activated pacemaker channel of brain. *Cell* **93,** 717–729 (1998).
2. Ludwig, A., Zong, X., Jeglitsch, M., Hofmann, F. & Biel, M. A family of hyperpolarization-activated mammalian cation channels. *Nature* **393,** 587–591 (1998).
3. Gauss, R., Seifert, R. & Kaupp, U. B. Molecular identification of a hyperpolarization-activated channel in sea urchin sperm. *Nature* **393,** 583–587 (1998).
4. Ishii, T. M., Takano, M., Xie, L. H., Noma, A. & Ohmori, H. Molecular characterization of the hyperpolarization-activated cation channel in rabbit heart sinoatrial node. *J. Biol. Chem.* **274,** 12835–12839 (1999).
5. Santoro, B. & Tibbs, G. R. The HCN gene family: molecular basis of the hyperpolarization-activated pacemaker channels. *Ann. N. Y. Acad. Sci.* **868,** 741–764 (1999).
6. Kaupp, U. & Seifert, R. Molecular diversity of pacemaker ion channels. *Annu. Rev. Physiol.* **63,** 235–257 (2001).
7. DiFrancesco, D. Pacemaker mechanisms in cardiac tissue. *Annu. Rev. Physiol.* **55,** 455–472 (1993).
8. Pape, H. C. Queer current and pacemaker: the hyperpolarization-activated cation current in neurons. *Annu. Rev. Physiol.* **58,** 299–327 (1996).
9. Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19,** 235–263 (1996).
10. Cui, J., Melman, Y., Palma, E., Fishman, G. I. & McDonald, T. V. Cyclic AMP regulates the HERG K^+ channel by dual pathways. *Curr. Biol.* **10,** 671–674 (2000).
11. DiFrancesco, D. & Tortora, P. Direct activation of cardiac pacemaker channels by intracellular cyclic AMP. *Nature* **351,** 145–147 (1991).
12. Santoro, B. *et al.* Molecular and functional heterogeneity of hyperpolarization-activated pacemaker channels in the mouse CNS. *J. Neurosci.* **20,** 5264–5275 (2000).
13. Weber, I. T. & Steitz, T. A. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. *J. Mol. Biol.* **198,** 311–326 (1987).
14. Goulding, E. H., Tibbs, G. R. & Siegelbaum, S. A. Molecular mechanism of cyclic-nucleotide-gated channel activation. *Nature* **372,** 369–374 (1994).
15. Varnum, M. D., Black, K. D. & Zagotta, W. N. Molecular mechanism for ligand discrimination of cyclic nucleotide-gated channels. *Neuron* **15,** 619–625 (1995).
16. Tibbs, G. R., Liu, D. T., Leypold, B. G. & Siegelbaum, S. A. A state-independent interaction between ligand and a conserved arginine residue in cyclic nucleotide-gated channels reveals a functional polarity of the cyclic nucleotide binding site. *J. Biol. Chem.* **273,** 4497–4505 (1998).
17. Barbuti, A., Baruscotti, M., Altomare, C., Moroni, A. & DiFrancesco, D. Action of internal pronase on the f-channel kinetics in the rabbit SA node. *J. Physiol.* **520,** 737–744 (1999).
18. Gordon, S. E. & Zagotta, W. N. Localization of regions affecting an allosteric transition in cyclic nucleotide-activated channels. *Neuron* **14,** 857–864 (1995).
19. Zong, X., Zucker, H., Hofmann, F. & Biel, M. Three amino acids in the C-linker are major determinants of gating in cyclic nucleotide-gated channels. *EMBO J.* **17,** 353–362 (1998).
20. Paoletti, P., Young, E. C. & Siegelbaum, S. A. C-Linker of cyclic nucleotide-gated channels controls coupling of ligand binding to channel gating. *J. Gen. Physiol.* **113,** 17–34 (1999).
21. Chen, J., Mitcheson, J. S., Lin, M. & Sanguinetti, M. C. Functional roles of charged residues in the putative voltage sensor of the HCN2 pacemaker channel. *J. Biol. Chem.* **275,** 36465–36471 (2000).
22. Chen, S., Wang, J. & Siegelbaum, S. A. Domains important for gating and cAMP regulation of HCN channels. *Soc. Neurosci. Abstr.* **26,** 2139 (2000).
23. Shin, K., Rothberg, B. & Yellen, G. Blocker state dependence and trapping in hyperpolarization-activated cation channels. Evidence for an intracellular activation gate. *J. Gen. Physiol.* **117,** 91–102 (2000).
24. del Camino, D., Holmgren, M., Liu, Y. & Yellen, G. Blocker protection in the pore of a voltage-gated K^+ channel and its structural implications. *Nature* **403,** 321–325 (2000).
25. Gulbis, J. M., Zhou, M., Mann, S. & MacKinnon, R. Structure of the cytoplasmic beta subunit-T1 assembly of voltage-dependent K^+ channels. *Science* **289,** 123–127 (2000).
26. Cushman, S. J. *et al.* Voltage dependent activation of potassium channels is coupled to T1 domain. *Nature Struct. Biol.* **7,** 403–407 (2000).
27. Kobertz, W. R., Williams, C. & Miller, C. Hanging gondola structure of the T1 domain in a voltage-gated K^+ channel. *Biochemistry* **39,** 10347–10352 (2000).
28. Liman, E. R., Tytgat, J. & Hess, P. Subunit stoichiometry of a mammalian K^+ channel determined by construction of multimeric cDNAs. *Neuron* **9,** 861–871 (1992).

© 2001 Macmillan Magazines Ltd

Acknowledgements

We thank J. Riley, H. Yao and E. Odell for their technical assistance. This work was partly supported by grants from the Whitehall Foundation (G.R.T.), NIH Medical Scientist Training Program (B.J.W.) and NINDS (S.A.S).

Correspondence and requests for materials should be addressed to G.R.T (e-mail: grt1@columbia.edu).

Sensitive detection of pathological prion protein by cyclic amplification of protein misfolding

Gabriela P. Saborio, Bruno Permanne & Claudio Soto

Serono Pharmaceutical Research Institute, CH1228 Geneva, Switzerland

Prions are the infectious agents responsible for transmissible spongiform encephalopathies. The principal component of prions is the glycoprotein PrP^{Sc}, which is a conformationally modified isoform of a normal cell-surface protein called PrP^{C} (ref. 1). During the time between infection and the appearance of the clinical symptoms, minute amounts of PrP^{Sc} replicate by conversion of host PrP^{C}, generating large amounts of PrP^{Sc} aggregates in the brains of diseased individuals. We aimed to reproduce this event *in vitro*. Here we report a procedure involving cyclic amplification of protein misfolding that allows a rapid conversion of large excess PrP^{C} into a protease-resistant, PrP^{Sc}-like form in the presence of minute quantities of PrP^{Sc} template. In this procedure, conceptually analogous to polymerase chain reaction cycling, aggregates formed when PrP^{Sc} is incubated with PrP^{C} are disrupted by sonication to generate multiple smaller units for the continued formation of new PrP^{Sc}. After cyclic amplification more than 97% of the protease-resistant PrP present in the sample corresponds to newly converted protein. The method could be applied to diagnose the presence of currently undetectable prion infectious agent in tissues and biological fluids, and may provide a unique opportunity to determine whether PrP^{Sc} replication results in the generation of infectivity *in vitro*.

To evaluate the cyclic amplification procedure (Fig. 1), we diluted brain homogenate from scrapie-affected hamsters until the signal of PrP^{Sc} was barely detectable by immunoblot after digestion with proteinase K (Fig. 2a, lane 1). The same dilution of scrapie brain homogenate was incubated with brain homogenate from healthy hamsters as a source of PrP^{C}. After incubation, an increase in the signal of a protease-resistant PrP^{Sc}-like protein (PrPres) with relative molecular mass 27,000–30,000 (M_r 27–30K) could be detected (Fig. 2a, lane 2). Using the same conditions, but subjecting the mixture to five cycles of incubation–sonication, the amount of PrPres dramatically increased (Fig. 2a, lane 3). This process of *in vitro* replication of protein conformation is called protein-misfolding cyclic amplification (PMCA). To estimate the rate of amplification, we performed densitometric analysis of immunoblots from five independent experiments done under the same conditions. The average amplification rate was 57.9 ± 19.9, indicating that after five PMCA cycles the newly converted protein corresponds to 97.4–98.7% of the total amount of PrPres. This proportion can be further increased by subjecting the sample to a larger number of amplification cycles (see below). To rule out artefacts of protein blotting, rat brain homogenate was added to the control sample of scrapie hamster brain homogenates to achieve equal protein concentrations (Fig. 2b). Only the hamster-derived protein was measured as the antibody used to reveal PrPres in the immunoblot does not react with rat prion protein. The conversion was dependent on the presence of PrP^{Sc}, as no PrPres formation was observed when the normal hamster brain homogenate was incubated alone, under the same conditions, with or without sonication (Fig. 2c, lanes 2 and 3). In addition, the increase in PrPres signal was not due to higher reactivity of initial PrP^{Sc} as a result of the treatment, because incubation–sonication of the material in the absence of a source of PrP^{C} resulted in no change in the signal (Fig. 2c, lanes 5 and 7). We noted that in the control experiment in which brain homogenates containing PrP^{C} and PrP^{Sc} were incubated, but not subjected to PMCA (Fig. 2a, lane 2), PrPres formation was greater than that reported previously using purified proteins[2,3] or cell lysates[4]. The higher efficiency of conversion might be due to the presence of additional factors present in the brain homogenate that catalyse the conversion[4,5]. The amplification system could be useful to isolate such additional factors by providing an assay to monitor their activity.

To evaluate the minimum PrP^{Sc} concentration needed to trigger amplification and enhanced detection of abnormal protein, we serially diluted brain homogenate from hamsters with scrapie directly into brain homogenate from healthy hamsters and assayed for PrPres signal with or without incubation–sonication cycles. Without PMCA, the PrPres signal diminished progressively until it

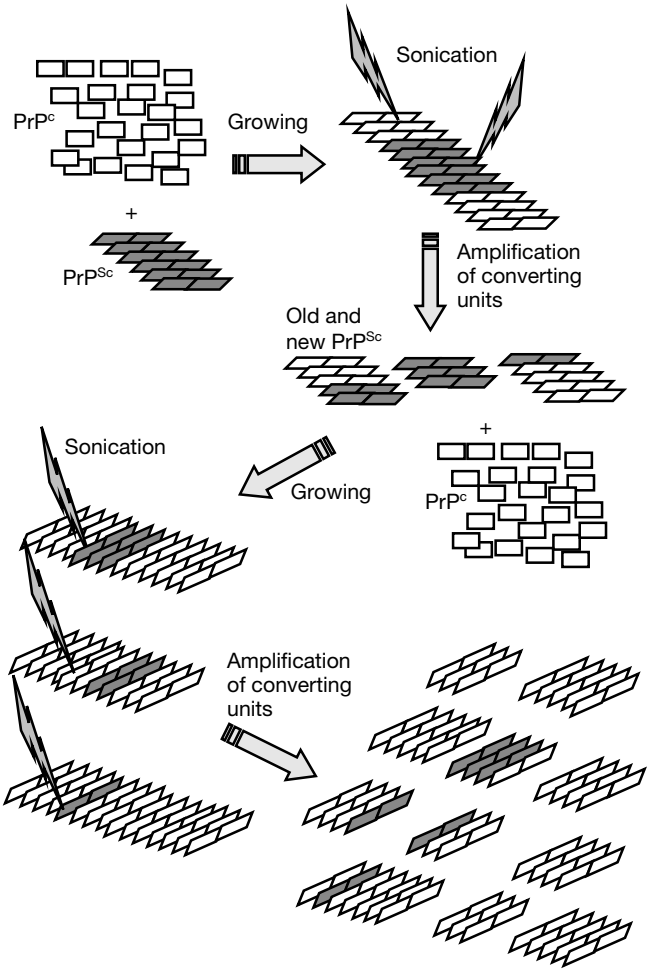


Figure 1 Diagrammatic representation of the PMCA procedure. Amplification is based on multiple cycles of PrP^{Sc} incubation in the presence of excess PrP^{C} followed by sonication. During the incubation periods, the size of oligomeric PrP^{Sc} is increased by incorporation of PrP^{C} into the growing aggregate, while during sonication the aggregates are disrupted, producing an expanded population of converting units.

© 2001 Macmillan Magazines Ltd

was not longer detectable at ~640-fold dilution (Fig. 3a, c). By contrast, when parallel samples were subjected to 10 PMCA cycles, PrPres was readily detected at this dilution (Fig. 3b, c). Indeed, PrPres was detected even after >10,000-fold dilution under these conditions. On the basis of our estimation of PrP^{Sc} detection by immunoblot of known amounts of recombinant hamster PrP, the minimum amount of PrP^{Sc} detectable under these conditions is ~6–12 pg or $0.2–0.4\times10^{-15}$ mol. The amount of newly converted PrPres is ~250 pg or 8.3×10^{-15} mol. This detection limit was calculated using low-sensitivity immunoblot assays and therefore it might be further decreased, to reach a similar or even better sensitivity than the biological assay of infectivity, by using detection systems with higher sensitivity and/or by performing a larger number of amplification cycles (see below).

To determine the effect of different numbers of PMCA cycles on PrPres formation, we diluted scrapie brain homogenate 1,000-fold and incubated it with an excess of healthy hamster brain homogenate. Samples were treated with 0, 5, 10, 20 or 40 cycles and the resultant PrPres signal determined by immunoblot. As anticipated, the levels of PrPres increased with the number of cycles (Fig. 4a). Control samples containing identical mixtures were incubated for the same time, but without sonication. No increase in PrPres signal strength was detected for these samples (Fig. 4b). The observed increase in PrPres per cycle fitted an exponential curve ($r^2 = 0.973$, Fig. 4c), limited by an underestimation of the amplified material at higher cycle numbers. Samples subjected to more than 20 cycles formed strong aggregates that could not be converted to monomeric protein during electrophoresis and hence were excluded from the quantification.

Propagation of transmissible spongiform encephalopathies (TSE) is believed to depend on the replication of PrP^{Sc} at the expense of the normal protein[1,6]. Although the molecular details of the replication process are not completely known, it involves changes in conformation that are stabilized upon protein oligomerization[6–8]. It is assumed that this process occurs *in vivo*, taking months or even years after infection of the host for PrP^{Sc} replication to progress sufficiently to trigger appearance of the disease. PrP replication has been performed *in vitro* in a cell-free system, by mixing purified PrP^{C} with an equimolar concentration of partially denatured PrP^{Sc} (ref. 2). This system has been used to study the molecular mechanism of PrP conversion[3], the sequence specificity of PrP^{Sc} formation[9], and to identify and evaluate inhibitors of PrP transformation[10,11]. However, the efficiency of the cell-free conversion is low, as the amount of newly converted protein is much less than the initial concentration of PrP^{Sc} needed to trigger the conversion reaction. This problem has precluded the study of the structural and infectious properties of the newly converted PrPres[6]. Our aim was to develop a system that mimics the replication of prions in the body during infection, by starting with undetectable concentrations of PrP^{Sc}, mixing it with a large excess

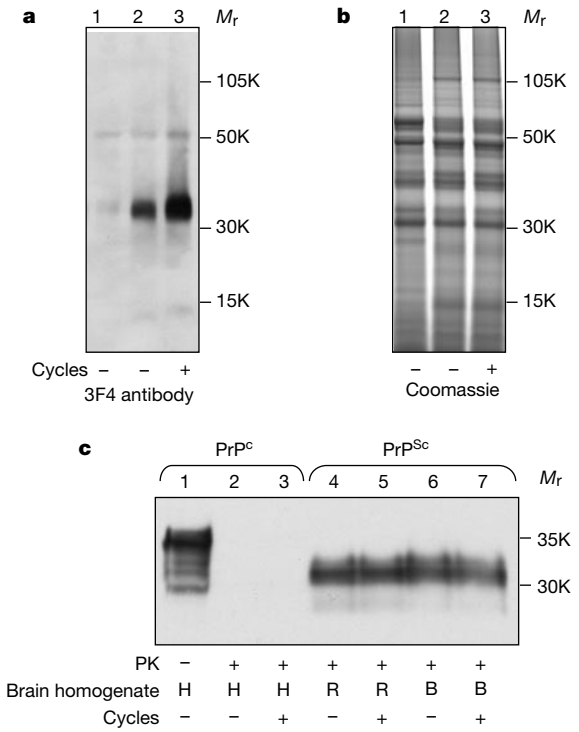


Figure 2 Amplification of PrP^{Sc} by sonication cycles. Scrapie hamster brain homogenate was diluted 500-fold into 10% healthy rat brain homogenate (lane 1 in **a** and **b**, control experiment) or into 10% healthy hamster brain homogenate (lanes 2 and 3 in **a** and **b**). Samples were incubated during 5 h without (lanes 1 and 2 in **a** and **b**) or with (lane 3 in **a** and **b**) five PMCA cycles. All samples were treated with proteinase K (PK). Samples were immunoblotted with the 3F4 anti-PrP antibody (**a**) or loaded onto a gel and stained for total protein with Coomassie blue (**b**). **c**, Control experiments with brain homogenate of healthy hamsters containing PrP^{C} (lanes 1–3) or brain homogenate of scrapie-affected hamsters containing PrP^{Sc} (lanes 4–7), were diluted 100-fold in healthy hamster brain homogenate (H), healthy rat brain homogenate (R) or PBS buffer (B), as indicated. Samples in lanes 1, 2, 4 and 6 were incubated for 5 h at 37 °C; samples in lanes 3, 5 and 7 were subjected to five PMCA cycles.

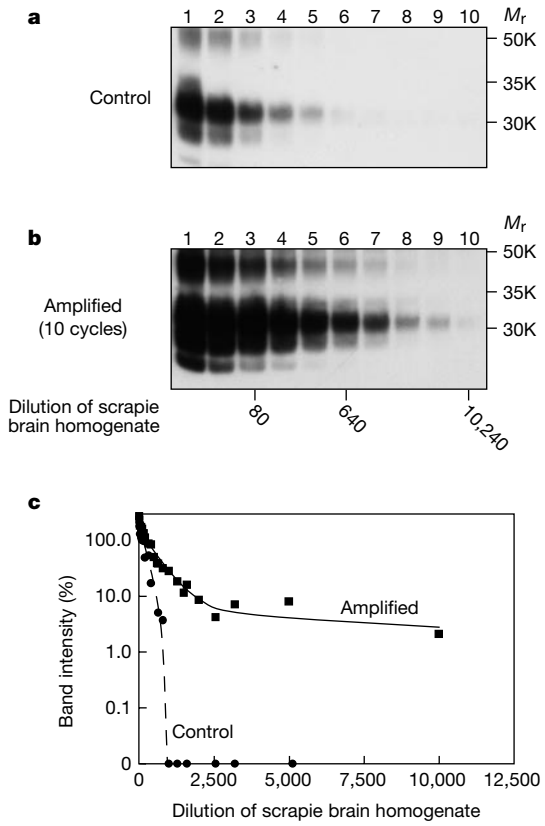


Figure 3 Sensitivity of the PMCA system. **a**, Scrapie hamster brain homogenate diluted serially into 5% rat brain homogenate (control). **b**, Two-fold serial dilutions of the scrapie brain homogenate into a 5% healthy hamster brain homogenate and subjected to 10 PMCA cycles. All samples in **a** and **b** were treated with proteinase K. The dilutions of scrapie brain homogenate in lanes 1–10 of **a** and **b** were 20, 40, 80, 160, 320, 640, 1,280, 2,560, 5,120 and 10,240, respectively. **c**, Densitometric quantification of five independent experiments to evaluate the sensitivity threshold of the cyclic amplification under these conditions. To normalize differences between experiments, the band intensity at each dilution is expressed as a percentage of the PrPres band intensity obtained at a 100-fold dilution of the scrapie brain homogenate without any treatment.

© 2001 Macmillan Magazines Ltd

of PrP^C, and finishing the reaction with most of the molecules in the altered conformation. Using the PMCA procedure we estimate that the initial amount of PrP^{Sc} corresponds to less than 3% of the total concentration of PrPres produced at the end of the conversion. The amplification procedure requires the presence of several factors: exogenous PrP^{Sc} acting as a template for the conversion; PrP^C that serves as a substrate; and unknown factors present in brain homogenate that catalyse the reaction (G.P.S. and C.S., unpublished observations). The strategy to speed up the conversion, and thus reproduce in a few hours *in vitro* a process that takes months or years for completion *in vivo*, was to perform cycles of incubation–sonication. It has been proposed that the infective unit of PrP^{Sc} is a β-sheet-rich oligomer that converts the normal protein by integrating it into the growing aggregate, where it acquires the properties associated with the abnormal protein[12,13]. After incubation of the two forms of PrP, the oligomeric species increase in size by recruiting and transforming PrP^C molecules. In the PMCA system, PrP^{Sc} is incubated with an excess of non-pathogenic conformer to enlarge the oligomeric converting units, followed by a sonication step to break down the aggregates into smaller units, each of which is capable of initiating further rounds of growth (Fig. 1). Our results demonstrate for the first time that the folding and biochemical properties of a protein can be transferred cyclically to other protein molecules, resulting in amplification of protein conformation in a manner conceptually analogous to DNA amplification by polymerase chain reaction (PCR).

A highly debated issue in the field of TSE is the nature of the infectious agent[14–16]. Strong evidence supports the 'protein-only' hypothesis of TSE propagation[1,6]. However, the conclusive proof of *de novo* production of prions is still lacking[6,15]. Our results show for the first time a high-efficiency replication of PrP conformation *in vitro*, mimicking the process of PrP^{Sc} replication thought to occur *in vivo* during the disease. Therefore, our findings reinforce one of the aspects of the prion hypothesis, that minute amounts of PrP^{Sc} have the ability to replicate its biochemical properties by converting large amounts of PrP^C. Cyclic amplification of PrP^{Sc} in the presence of PrP^C provides a unique opportunity to generate infectivity *in vitro*. The protein generated with the PMCA procedure shares several of the biochemical properties typical of PrP^{Sc} extracted from the brains of animals with scrapie, including protease-resistance, detergent insolubility (data not shown) and the ability to further convert PrP^C *in vitro*. Because of the high yield of conversion, the infectious properties of the newly generated PrP^{Sc}-like protein can be tested and distinguished from the small amount of PrP^{Sc} used to begin the conversion reaction. We are currently determining the infectious properties of the PrP^{Sc}-like protein generated by PMCA.

The definitive diagnosis of TSE is based on the demonstration of PrP^{Sc} in brain tissue of the affected host[17]. Indeed, PrP^{Sc} is the only validated surrogate marker for the disease and the only known component of the TSE infectious agent[1]. The lack of a prion-specific nucleic acid prevents the use of highly sensitive PCR-based diagnostic tests. Given the recent widespread distribution of cases of bovine spongiform encephalopathy (BSE) in Europe and the strong arguments linking it to variant Creutzfeldt–Jakob disease (CJD) in humans[18,19], development of a sensitive diagnostic test that can reliably identify the disease in animals and people during the presymptomatic period is a top priority[20]. Currently available methods to detect PrP^{Sc} are limited by the amount of the abnormal protein

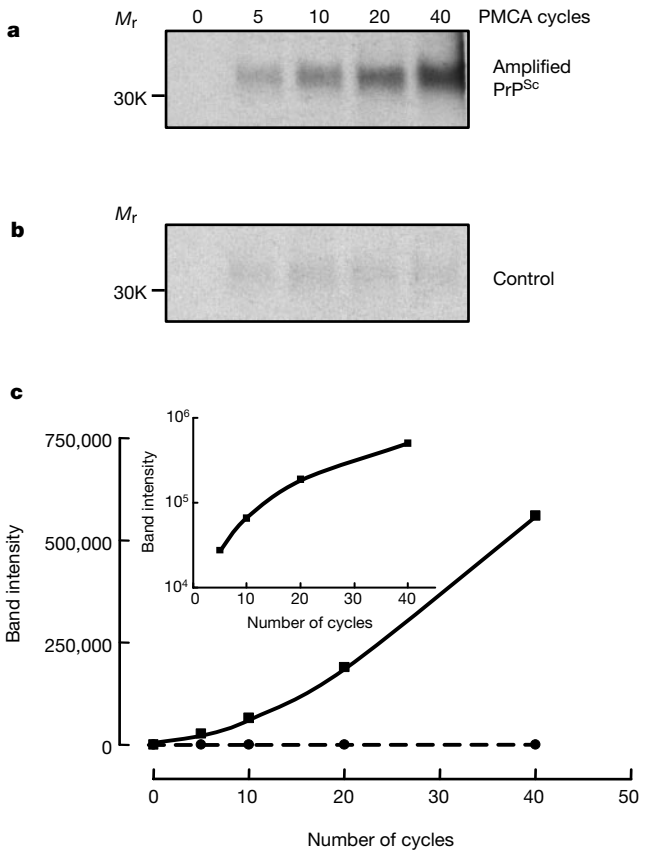


Figure 4 Relationship between the extent of the conversion and the number of amplification cycles. A 1,000-fold dilution of hamster scrapie brain homogenate was made into 5% healthy hamster brain homogenate. Aliquots were subjected to 0, 5, 10, 20 or 40 PMCA cycles (**a**) or incubated for the same times, but without sonication (**b**). All samples were treated with proteinase K and PrPres signal was evaluated by immunoblot. To appreciate more clearly the differences in the band intensity, one tenth of the samples were loaded in the gel. **c**, Densitometric quantification of the blots in **a** (PMCA, solid line) and **b** (control, dotted line); the inset shows the same data in a logarithmic scale.

© 2001 Macmillan Magazines Ltd

present in the tissues[21]. This limitation results in late-stage diagnosis, in most cases from brain tissue obtained at autopsy[20,21]. However, as shown by infectivity studies, the infectious agent is present pre-symptomatically in many tissues in addition to the brain[21,22]. Thus, early detection of PrP^{Sc} from non-brain sources should be possible with highly sensitive methods[21]. Our findings constitute a strategy to detect low quantities of PrP^{Sc}, by means of amplifying undetectable amounts of the protein to a detectable level. The PMCA procedure can be combined with any of the existing detection systems to reach a further reduction of detection threshold. Preliminary results indicate that the PMCA system can also be applied to human brain samples obtained from sporadic CJD (G.P.S. and C.S., unpublished observations). Therefore, the PMCA method opens a new possibility for TSE diagnosis that could be applied to systemic tissues or fluids during the pre-symptomatic phase of the disease. □

Methods

Preparation of brain homogenates

Brains from healthy rats (F.344) and Syrian golden hamsters healthy or infected with the adapted scrapie strain 263 K were obtained after decapitation and immediately frozen in dry ice and kept at −80 °C until used. Brains were homogenized in PBS buffer containing protease inhibitors (Complete cocktail from Boehringer Mannheim) at a 1× final concentration. Detergents (0.5% Triton X-100, 0.05% SDS, final concentrations) were added and samples clarified with low-speed centrifugation (1,000*g*) for 1 min, using an Eppendorf 5415 centrifuge.

Cyclic amplification

Serial dilutions of the scrapie brain homogenate were made directly in the healthy brain homogenate. We incubated 60 μl of these dilutions at 37 °C with agitation. Every hour one cycle of sonication (five pulses of 1 s each) was done using a microsonicator (Bandelin Electronic, Sonopuls) with the probe immersed in the sample and the power setting fixed at 40%. These cycles were repeated 5–40 times.

Detection of PrPres

The samples were digested with proteinase K (100 μg ml^{-1} for 60 min at 37 °C) and the reaction was stopped with 50 mM phenyl-methyl sulphonyl fluoride. Samples were separated by SDS–PAGE and electroblotted into nitrocellulose membrane in 3-(cyclohexylamino)-1-propane sulphonic acid or Tris-glycine transfer buffer with 10% methanol during 45 min at 400 mA. For immunoblotting, the membranes were blocked with 5% non-fat milk and incubated for 2 h with the monoclonal antibody 3F4 (ref. 23) (1:50,000). Four washes of 5 min each were performed with PBS and 0.3% Tween20 before incubation with secondary anti-mouse antibody labelled with horseradish peroxidase (1:5,000) for 1 h. After washing, the reactivity in the membrane was developed with an ECL Chemilumicense Kit (Amersham) according to the manufacturer's instructions. Densitometric analyses of western blots were performed with the program SigmaGel v1.0 (Jandel Scientific). The concentration of PrP^{Sc} was estimated by densitometric analysis and comparison with pure recombinant hamster PrP, the concentration of which was determined by amino-acid analysis.

Received 13 February; accepted 10 May 2001.

1. Prusiner, S. B. Prions. *Proc. Natl Acad. Sci. USA* **95,** 13363–13383 (1998).
2. Kocisko, D. A. *et al.* Cell-free formation of protease-resistant prion protein. *Nature* **370,** 471–474 (1994).
3. Horiuchi, M. & Caughey, B. Prion protein interconversions and the transmissible spongiform encephalopathies. *Structure Fold Des.* **7,** R231–R240 (1999).
4. Saborio, G. P. *et al.* Cell-lysate conversion of prion protein into its protease-resistant isoform suggests the participation of a cellular chaperone. *Biochem. Biophys. Res. Commun.* **258,** 470–475 (1999).
5. Telling, G. C. *et al.* Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell* **83,** 79–90 (1995).
6. Aguzzi, A. & Weissmann, C. Prion research: the next frontier. *Nature* **389,** 795–798 (1997).
7. Cohen, F. E. & Prusiner, S. B. Pathologic conformations of prion proteins. *Ann. Rev. Biochem.* **67,** 793–819 (1998).
8. Caughey, B., Kocisko, D. A., Raymond, G. J. & Lansbury, P. T. Jr Aggregates of scrapie-associated prion protein induce the cell-free conversion of protease-sensitive prion protein to the protease-resistance state. *Chem. Biol.* **2,** 807–817 (1995).
9. Kocisko, D. A. *et al.* Species specificity in the cell-free conversion of prion protein to protease-resistant forms: a model for the scrapie species barrier. *Proc. Natl Acad. Sci. USA* **92,** 3923–3927 (1995).
10. Chabry, J., Caughey, B. & Chesebro, B. Specific inhibition of *in vitro* formation of protease-resistant prion protein by synthetic peptides. *J. Biol. Chem.* **273,** 13203–13207 (1998).
11. Caughey, W. S., Raymond, L. D., Horiuchi, M. & Caughey, B. Inhibition of protease-resistant prion protein formation by porphyrins and phthalocyanines. *Proc. Natl Acad. Sci. USA* **95,** 12117–12122 (1998).
12. Harper, J. D. & Lansbury, P. T. Jr Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Ann. Rev. Biochem.* **66,** 385–407 (1997).
13. Brown, P., Goldfarb, L. G. & Gajdusek, D. C. The new biology of spongiform encephalopathy: infectious amyloidoses with a genetic twist. *Lancet* **337,** 1019–1022 (1991).
14. Chesebro, B. BSE and prions: uncertainties about the agent. *Science* **279,** 42–43 (1998).
15. Mestel, R. Putting prions to the test. *Science* **273,** 184–189 (1996).
16. Dormont, D. Agents that cause transmissible subacute spongiform encephalopathies. *Biomed. Pharmacother.* **53,** 3–8 (1999).
17. Johnson, R. T. & Gibbs, C. J. Creutzfeldt–Jakob disease and related transmissible spongiform encephalopathies. *New Engl. J. Med.* **339,** 1994–2004 (1998).
18. Collinge, J. Variant Creutzfeldt–Jakob disease. *Lancet* **354,** 317–323 (1999).
19. Bruce, M. E. *et al.* Transmissions to mice indicate that new variant CJD is caused by the BSE agent. *Nature* **389,** 498–501 (1997).
20. Schiermeier, Q. Testing times for BSE. *Nature* **409,** 658–659 (2001).
21. Brown, P., Cervenakova, L. & Diringer, H. Blood infectivity and the prospects for a diagnostic screening test in Creutzfeldt–Jakob disease. *J. Lab. Clin. Invest.* **137,** 5–13 (2001).
22. Glatzel, M. & Aguzzi, A. Peripheral pathogenesis of prion diseases. *J. Gen. Virol.* **81,** 2813–2821 (2000).
23. Kascsak, R. J. *et al.* Mouse polyclonal and monoclonal antibody to scrapie-associated fibril proteins. *J. Virol.* **61,** 3688–3693 (1987).

Acknowledgements

We thank S. Peano, A. Conz, L. Anderes and M.-J. Frossard for technical assistance and R. J. Kascsak for providing 3F4 anti-PrP antibody. We are grateful to M. Pocchiari, S. Fumero, T. Wells, K. Maundrell and J. DeLamarter for reading the manuscript and providing comments. We also thank C. Herbert for help in the preparation of the figures.

Correspondence and requests for materials should be addressed to C.S. (e-mail: claudio.soto@serono.com) or to G.P.S. (e-mail: gabriela.saborio@serono.com).

Quenching quorum-sensing-dependent bacterial infection by an *N*-acyl homoserine lactonase

Yi-Hu Dong*, Lian-Hui Wang*, Jin-Ling Xu, Hai-Bao Zhang, Xi-Fen Zhang & Lian-Hui Zhang

Laboratory of Biosignals and Bioengineering, Institute of Molecular Agrobiology, National University of Singapore, 1 Research Link, 117604, Singapore
** These authors contributed equally to this work*

Bacterial cells sense their population density through a sophisticated cell–cell communication system and trigger expression of particular genes when the density reaches a threshold. This type of gene regulation, which controls diverse biological functions including virulence, is known as quorum sensing[1,2]. Quorum-sensing signals, such as acyl-homoserine lactones (AHLs), are the essential components of the communication system. AHLs regulate virulence gene expression in a range of plant and animal (including human) bacterial pathogens[3–9]. AHL-producing tobacco restored the pathogenicity of an AHL-negative mutant of *Erwinia carotovora*[10]. Different bacterial species may produce different AHLs, which vary in the length and substitution of the acyl chain but contain the same homoserine lactone moiety. Here we show that the acyl-homoserine lactonase (AHL-lactonase), a new enzyme from *Bacillus* sp.[11], inactivates AHL activity by hydrolysing the lactone bond of AHLs. Plants expressing AHL-lactonase quenched pathogen quorum-sensing signalling and showed significantly enhanced resistance to *E. carotovora* infection. Our results highlight a promising potential to use quorum-sensing signals as molecular targets for disease control, thereby broadening current approaches for prevention of bacterial infections.

Target genes regulated by AHLs are extremely varied and regulatory mechanisms are probably diversified[12–14]; however, the general mechanism of AHL-mediated quorum-sensing signalling is highly conserved. In general, each bacterial cell produces a basal level of AHLs that move in and out of cell membranes through diffusion or active transportation (ref. 15 and L.-H.Z., unpublished data). When AHLs reach a threshold concentration owing to

© 2001 Macmillan Magazines Ltd

higher bacterial population density, they interact with a LuxR family transcription factor, and initiate expression of particular genes[13,14,16]. Recently, we showed that the *aiiA* gene from a Gram-positive *Bacillus* sp. 240B1 encodes an enzyme capable of inactivating the three AHLs tested[11]; also a bacterial isolate of *Variovorax paradoxus* has been identified that degrades AHLs and releases homoserine lactone, indicating the presence of an aminoacylase[17]. These findings suggest that it could be possible to establish a generic 'quorum-quenching' approach to control bacterial infections, that is, to paralyse quorum-sensing systems of bacterial pathogens through inactivation of quorum-sensing signals.

To determine how the enzyme encoded by *aiiA* inactivates AHL signals, we first tested *N*-(3-oxohexanoyl)-L-homoserine lactone (OHHL), which regulates expression of virulence genes in a plant pathogen *Erwinia carotovora* pv. *carotovora*[3,5]. We digested OHHL with the enzyme and analysed the reaction products with high-performance liquid chromatography (HPLC) and mass spectrometry. Enzymatic digestion of OHHL resulted in only one product fraction, which was more hydrophilic than OHHL, as determined by HPLC analysis. Electrospray ionization mass spectrometry (ESI-MS) analysis of this product showed a strong quasimolecular (M-H) ion at an *m*/*z* (mass-to-charge ratio) of 230.1 (Fig. 1a; left column, top), suggesting that the enzymatic action on OHHL (M-H ion *m*/*z* of 212) leads to a mass increase of 18 in its product, corresponding to a water molecule. This is in agreement with the chemical composition of the lactone-opened OHHL, namely *N*-(3-oxohexanoyl)-L-homoserine (M-H ion *m*/*z* of 230). Tandem mass spectrometry of the parent ion at *m*/*z* of 230 (MS^2) showed a daughter ion of 118.1 (Fig. 1a; left column, middle), consistent with the formula of $C_4H_9NO_3$ (homoserine, M-H ion *m*/*z* of 118.1). Further analysis of the ion of *m*/*z* of 118.1 (MS^3) identified its daughter ion of *m*/*z* of 100 (Fig. 1a; left column, bottom). Homoserine produced a parent ion with a *m*/*z* of 118.1 and a daughter ion with a *m*/*z* of 100 (data not shown). To further confirm the analysis, we produced *N*-(3-oxohexanoyl)-L-homoserine from its corresponding lactone and found that its mass spectrum was identical to that of the hydrolysis product of OHHL, showing a parent ion *m*/*z* of 230 in ESI-MS and daughter ions at *m*/*z* of 118 and 100 in tandem MS (Fig. 1a, right column). HPLC analysis showed that the synthetic *N*-(3-oxohexanoyl)-L-homoserine has an identical HPLC retention time as the OHHL hydrolytic product.

We next analysed the enzyme-digested products of three other AHLs with differences in acyl-chain length and substitution at the C3 position, including *N*-(3-oxododecanoyl)-L-homoserine lactone (OdDHL), *N*-butanoyl-L-homoserine lactone (BHL), and *N*-(3-oxooctanoyl)-L-homoserine lactone (OOHL). OdDHL and BHL regulate production of virulence determinants in *Pseudomonas aeruginosa*[18,19], a human pathogen causing fatal infections of the lung and of burns. OOHL controls Ti plasmid conjugal transfer in *Agrobacterium tumefaciens*[20]. Results showed again that each of their molecular masses had been increased by 18 after the enzyme reaction (data not shown). In all cases, controls lacking the enzyme showed negligible delactonization. The results demonstrate that the enzyme encoded by *aiiA*, designated previously as AiiA[11], is an AHL-lactonase that hydrolyses the ester bond of the homoserine lactone ring of AHLs (Fig. 1b). The minimum concentrations for detectable biological activity of delactonized products of OHHL and OdDHL were 500 μM and 1000 μM, respectively, indicating a 1,667 times decrease in activity in comparison with OHHL and OdDHL. The residue activity of these delactonized products is unlikely to be due to spontaneous relactonization or formation of methyl ester derivatives, as HPLC and ESI-MS analysis did not detect the corresponding fractions even after incubation for two days with or without bioassay strain.

To test the effect of AHL-lactonase on bacterial infection, we introduced *aiiA* and an *sp*–*aiiA* fusion gene (see Methods), carried by constructs pBI-AI and pBI-SPAI, respectively (Fig. 2a), into the genome of tobacco and potato by means of *Agrobacterium*-mediated transformation[21,22]. The *sp*–*aiiA* fusion gene targets AHL-lactonase to the intercellular space of plant tissues where *E. carotovora* initiates infection. We randomly selected 38 kanamycin-resistant primary transgenic lines of tobacco (20 with *aiiA* and 18 with *sp*–*aiiA*) and 35 lines of potato (29 with *aiiA* and 6 with *sp*–*aiiA*) for further analysis, along with 10 controls consisting of 5 untransformed and 5 pBI-GFP (construct with green fluorescent protein coding region) transformed tobacco and potato plants.

RNA hybridization assays showed that the *aiiA* transcript was detected in all *aiiA* trangenic plants but not in control lines (Fig. 2b). Soluble protein samples extracted from transgenic tobacco leaves and potato tubers inactivated OHHL activity *in vitro* (Fig. 3a). Arbitrary enzyme activities of 7 tobacco and 5 potato lines ranged from 1 to 6 picomoles of OHHL hydrolysed by soluble protein

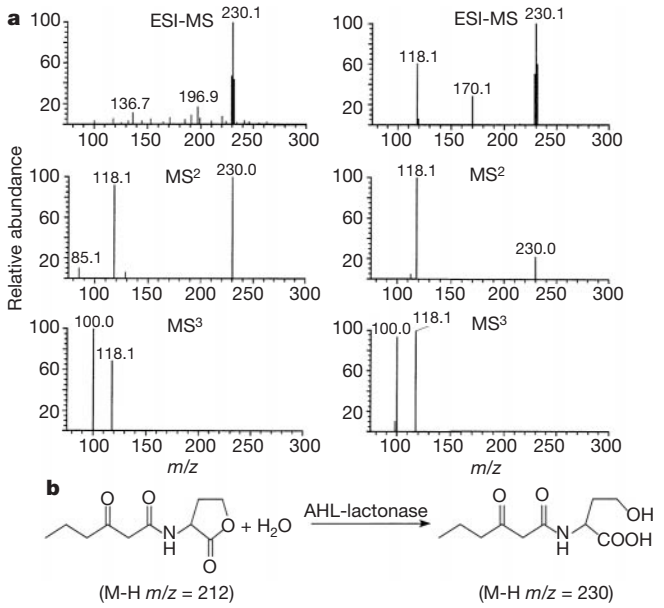


Figure 1 ESI-MS and tandem mass spectrometry analysis of the hydrolysis product of OHHL. **a**, Electrospray ionization spectra of the AHL-lactonase hydrolysis product of OHHL (left column) and of the produced *N*-(3-oxohexanoyl)-L-homoserine (right column). We labelled the comparable peaks with their respective *m*/*z*. **b**, Mechanism of AHL-lactonase in inactivating OHHL. AHL-lactonase opened the homoserine lactone ring of OHHL in the presence of water to produce *N*-(3-oxohexanoyl)-L-homoserine.

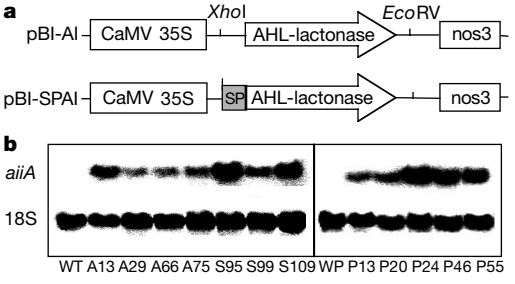


Figure 2 Genetic constructs for plant transformation and expression of *aiiA* messenger RNA in transgenic plants. **a**, Expression cassettes containing *aiiA* gene (pBI-AI) and the *sp*–*aiiA* fusion gene (pBI-SPAI). SP, signal peptide from the tobacco calreticulin; CMV35S, cauliflower mosaic virus 35S promoter; nos3, nopaline synthase transcription terminator. **b**, RNA gel blot analyses of *aiiA* in transgenic tobacco (left) and potato (right). We used a tobacco 18S rRNA (18S) probe to show relative RNA loadings. The A- and S-lines are tobacco *aiiA* and *sp*–*aiiA* transformants respectively; P-lines are potato *aiiA* transformants. WT, wild-type tobacco; WP, wild-type potato.

© 2001 Macmillan Magazines Ltd

($h^{-1}\,\mu g^{-1}$). The AHL-lactonase activities in plants transformed with the *sp–aiiA* fusion gene were generally lower than in those transformed with the *aiiA* gene.

Figure 3b shows the areas of maceration of control plants and transgenic lines of tobacco and potato after inoculation with *E. carotovora*. Enhanced resistance of the *aiiA*/*sp–aiiA* transgenic plants is consistent with the ability of AHL-lactonase to inactivate OHHL (Fig. 3a). In transgenic lines with high levels of AHL-lactonase protein (Fig. 3c), for example, tobacco line A13 and potato lines P13, P24 and P46, the areas of maceration are less than 5% of those in untransformed controls (Fig. 3b), indicating a strong correlation between disease resistance and AHL-lactonase expression level. Transcriptional expression levels in the three *sp–aiiA* tobacco lines were higher than in the *aiiA* transgenic lines (Fig. 2b), but their OHHL-hydrolysing enzyme activities and resistance levels were not as strong as in the best *aiiA* transgenic lines (Fig. 3a, b). The enzyme activity of the purified SP-AHL-lactonase from *Escherichia coli* was comparable to that of AHL-lactonase. It is possible that the reduced levels of delactonase activity in the *sp–aiiA* trangenic plants resulted from trapping the SP-AHL-lactonase in microsomes, as in the case of calreticulin protein[23], which could slow down the speed of contact between the AHL-lactonase and OHHL.

Immunoblot analysis confirmed the translational expression of the AHL-lactonase protein (Fig. 3c). Lines with higher AHL-lactonase enzyme activities (for example, A13, P13, P24 and P46) also showed stronger immunoblot signals than the lines with the lower enzyme activities. By using dilutions of purified AHL-lactonase as standards in immunoblot analysis, we estimated the AHL-lactonase protein contents to be 2–7 ng mg^{-1} and 20–110 ng mg^{-1} of soluble proteins in tobacco leaves and in potato tubers, respectively.

We found that resistance of the *aiiA* transgenic plants is related to the population density of the inoculated pathogen. Control plants including untransformed and GFP transgenic tobacco displayed typical maceration symptoms several hours after inoculation. The higher the population density of the inoculum, the larger the maceration area on the tobacco leaves (Fig. 4a). In the four tested *aiiA* transgenic lines, a low density of inoculum (600 colony-forming units (CFU)) did not cause any symptoms 20 h after inoculation. Maceration occurred with these plants when they were inoculated with a high population density of pathogen (60,000 CFU), but the macerated areas were much smaller than in the controls. We obtained similar results with normal and transgenic potato lines when challenged with the pathogen at varied cell densities (25,000–100,000 CFU) (Fig. 4c).

Even when maceration was initiated in the *aiiA* transgenic lines after inoculation with a high population density inoculum, further symptom development was significantly retarded or even stopped. Forty hours after inoculation, the leaves of control tobacco were totally macerated, whereas symptoms on the *aiiA* tobacco leaves were almost the same as at 20 h after inoculation (Fig. 4a, b). The effect of AHL-lactonase on symptom development appears even

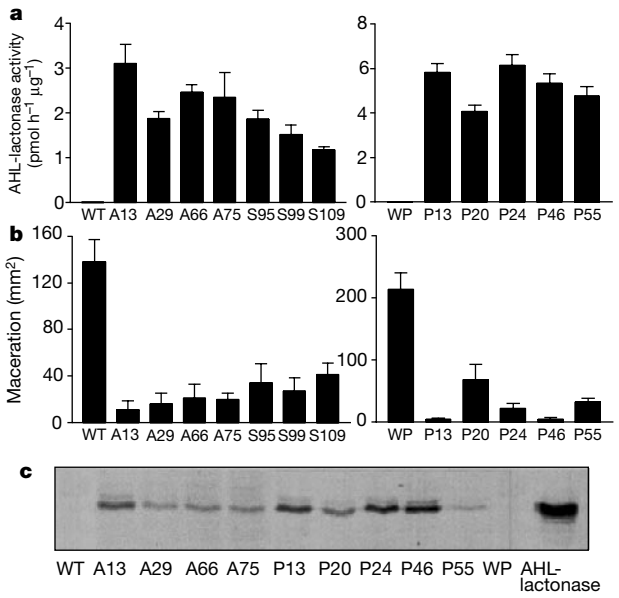


Figure 3 AHL-lactonase enzyme activity and *E. carotovora* inoculation. **a**, AHL-lactonase enzyme activity in transgenic tobacco (left) and potato (right). The line designation is the same as in Fig. 2. The data are means of 5 (tobacco) or 8 (potato) replicates. Vertical bar represents standard error. **b**, *Erwinia carotovora* SCG1 inoculation assay. The population density per inoculation site was 6×10^3 CFU for tobacco or 1×10^5 CFU for potato. The data are means of 8 (for tobacco) or 6 (for potato) replicates. **c**, Immunoblot analysis of AHL-lactonase protein in transgenic plants.

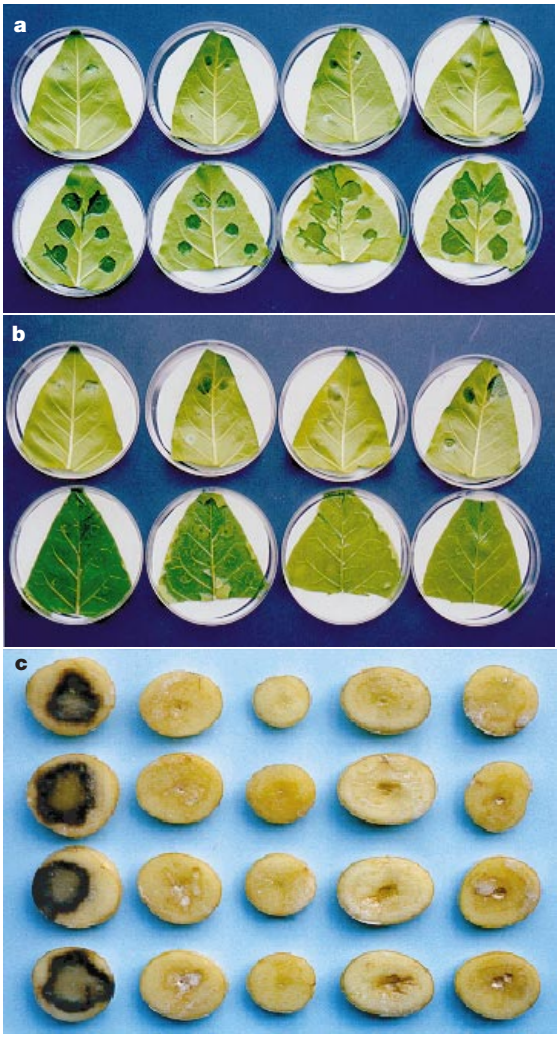


Figure 4 Plant inoculation with *Erwinia carotovora* SCG1. **a**, Top row are tobacco leaves from *aiiA* lines A13, A29, A66 and A75; bottom row are those from untransformed lines W6, W7, W9 and a GFP-line G7, respectively (from left to right). We inoculated each leaf on 6 sites following the same order. The inoculum cell numbers on the two top spots, the two spots on the lower left side, and the two spots on the lower right side of each leaf were 60,000, 6,000 and 600 CFU, respectively. We took the photograph 20 h after inoculation. **b**, Same as **a** except that we took the photograph 40 h after inoculation. **c**, The potato tuber slices in columns from left to right are from wild-type plants and from *aiiA* lines P46, P24, P55 and P13, respectively. The inoculum cell numbers for rows from top to bottom were 25,000, 50,000, 75,000 and 100,000 CFU, respectively. We took the photograph 48 h after inoculation.

© 2001 Macmillan Magazines Ltd

clearer in the *aiiA* transgenic potato. We observed watery rotten lesions in the control several hours after inoculation. Symptoms worsened progressively until the whole slice was completely macerated. In contrast, inoculation at high population density initially produced watery lesions in the tubers of *aiiA* potato but soon the lesions became dry and symptom development stopped (Fig. 4c). The probable pattern is that the AHL-lactonase by slowing down the pathogen from producing virulence factors, which include pectolytic enzymes, allows the host defense mechanisms to take over and eventually stop infection.

Our results demonstrate that enzymatic quenching of AHL quorum-sensing signals is a feasible approach for prevention of bacterial infection. Besides the two AHL-inactivation enzymes[11,17], two groups of AHL antagonists have also been documented. The halogenated furanones from marine red alga *Delisea pulchra* inhibited luminescence and virulence in *Vibrio harveryi* by displacement of OHHL from LuxR transcription factor[24–26]. The antimicrobial triclosan suppressed AHL biosynthesis through inhibiting the reaction catalysed by enoyl-acyl carrier protein reductase[27]. As the quorum-sensing regulation of virulence seems to be one of the common strategies that many bacterial pathogens have adopted during evolution to ensure their survival in host–pathogen interactions, this kind of quorum-quenching strategy could have wide applications in our fight against bacterial plagues. □

Methods

AHL-lactonase purification and structural determination

We purified the recombinant AHL-lactonase by using the glutathione *S*-transferase (GST) Gene Fusion System (Pharmacia) as described[11]. Briefly, the GST/AHL-lactonase fusion protein from *E. coli* cell-free extracts was bonded to glutathione affinity resins, and then recombinant AHL-lactonase was released from the bound GST protein by thrombin, a site-specific protease. Purity was determined by SDS–PAGE analysis. We mixed the purified AHL-lactonase (30 nmol) with OHHL (500 nmol) per milliliter of phosphate buffer (1/15 M, pH 8.0), and incubated the mixture at 28 °C for 10 min with gentle shaking. After incubation, we extracted the mixture three times with ethyl acetate, and evaporated the combined organic phase through a rotary evaporator. For HPLC analysis and purification, a sample was dissolved in 0.2 ml methanol and analysed with a symmetry C_{18} reverse-phase column (4.6 × 250 mm). Fractions were separated by eluting isocratically with 50:50 methanol/water (v/v) at a flow rate of 1 ml min^{-1}. ESI-MS and tandem mass spectrometry were performed on a Finnigan/MAT LCQ ion-trap mass spectrometer. The sample dissolved in 50:50 methanol/water (v/v) was introduced into the mass spectrometer by loop injection. We synthesized all the AHLs used in this study as previously described[20]. We produced *N*-(3-oxohexanoyl)-L-homoserine by incubating the corresponding acyl-homoserine lactone (50 mg) in 350 μl dimethyl sulphoxide containing 350 μl of 1 M NaOH for 12 h at 25 °C. After incubation, we adjusted the mixture to pH 6.5 with 1 M NaH_2PO_4 and then extracted it three times with chloroform. The chloroform fractions were combined and evaporated to dryness. The product was purified by HPLC using a C_{18} reverse-phase column. Its structure was confirmed by ^1H NMR and mass spectrometry.

Plant transformation and analysis

We cloned *aiiA* in a plant expression vector pBI121 that contains a kanamycin resistance gene as the selection marker and resulted in the construct pBI-AI (Fig. 2a; top panel). To test whether targeting of the AHL-lactonase enzyme into intercellular spaces could confer better resistance to *E. carotovora* infection, we prepared the second construct pBI-SPAI (Fig. 2a, bottom panel) by fusing the coding sequence of a 27-amino-acid calreticulin secretion signal peptide[23,28] to the 5′ end of the *aiiA* coding sequence in the same open reading frame. We also cloned the coding region of the GFP in pBI121 to produce the construct pBI-GFP as a control. These constructs are under the control of the cauliflower mosaic virus (CMV) 35S promoter. We introduced the pBI-AI, pBI-SPAI and pBI-GFP constructs into the genome of tobacco (*Nicotiana tabacum* L. var. GX3) and potato (*Solanum tuberosum* L. var. Bintje) plants respectively through *Agrobacterium*-mediated transformation[16,17]. Transgenic plants were grown under standard conditions (25 °C, 16/8 h light/dark cycle) in growth chambers.

We extracted total RNA from tobacco and potato leaves by the Trizol method (Gibco). Total RNA (10 μg) was fractionated in 1.2% formaldehyde agrose gel and transferred to Hybond–N^+ (Amersham) nylon membranes. The *aiiA* coding region and the 18S ribosomal RNA fragment of tobacco were labelled separately with α-P^{32}-deoxycytidine 5′-triphosphate (α-P^{32}-dCTP) using Rediprime II (Amersham). Hybridization was performed according to standard procedures.

We determined AHL-lactonase enzyme activity by incubation of OHHL with total soluble protein samples extracted from plants. Fresh tobacco leaf tissue or potato tuber tissue (1 g) was homogenized with 1.5 ml (for tobacco) or 1.0 ml (for potato) of protein extraction buffer (100 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 18% (w/v) sucrose, 40 mM 2-mercaptoethanol). Reaction mixture containing protein extracts and OHHL (40 μM) was incubated for 30 min at 28 °C. After digestion, we determined the remaining OHHL as previously described[11] and calculated the digested OHHL. AHL-lactonase activity is defined as picomoles of OHHL hydrolysed per hour per microgram of total soluble protein.

We prepared polyclonal anti-AHL-lactonase rabbit antiserum by multiple injections of the purified AHL-lactonase protein into rabbits. For immunoblot analysis, we separated soluble proteins from wild-type and transgenic plants as well as the purified AHL-lactonase protein (10 ng) by SDS–PAGE and visualized with anti-AHL-lactonase antibodies and a standard alkaline phosphatase immunoassay.

Plant inoculation

We prepared *E. carotovora* pv. *carotovora* SCG1 cells by centrifugation (5,000*g* for 10 min) of overnight culture and dilution to different population densities in phosphate buffer (1/15 M, pH 8.0). We excised tobacco leaves of the same age from 5–8-week-old plants and placed them in Petri dishes that contained one layer of Whatman filter paper moistened with 2 ml of sterile water. We inoculated the leaves by adding 3 μl SCG1 dilution on the freshly wounded leaf surface (punched slightly with a 20-μl pipette tip). For potato, we surface-sterilized mature tubers with 70% ethanol, sliced them 3 mm apart, and placed them on Petri dishes with wet filter paper to keep them moist. To each slice we added 5 μl SCG1 dilution after gentle tip punching, and incubated them at 28 °C for 20 h (tobacco) or 48 h (potato), unless otherwise indicated, before determining the extent of maceration.

Received 24 January; accepted 8 May 2001.

1. Fuqua, C., Winans, S. C. & Greenberg, E. P. Census and consensus in bacterial ecosystems: the LuxR-LuxI family of quorum-sensing transcriptional regulators. *Annu. Rev. Microbiol.* **50,** 727–751 (1996).
2. Robson, N. D., Cox, A. R., McGowan, S. J., Bycroft, B. W. & Salmond, G. P. Bacterial *N*-acyl-homoserine-lactone-dependent signaling and its potential biotechnological applications. *Trends Biotechnol.* **15,** 458–464 (1997).
3. Jones, S. M. *et al.* The *Lux* autoinducer regulates the production of exoenzyme virulence determination in *Erwinia carotovora* and *Pseudomonas aeruginosa*. *EMBO J.* **12,** 2477–2482 (1993).
4. Passador, L., Cook, J. M., Gambello, M. J., Rust, L. & Iglewski, B. H. Expression of *Pseudomonas aeruginosa* virulence genes requires cell-to-cell communication. *Science* **260,** 1127–1130 (1993).
5. Pirhonen, M., Flego, D., Heikinheimo, R., & Palva, E. A small diffusible signal molecule is responsible for the global control of virulence and exoenzyme production in the plant pathogen *Erwinia carotovora*. *EMBO J.* **12,** 2467–2476 (1993).
6. von Bodman, S. B. & Farrand, S. K. Capsular polysaccharide biosynthesis and pathogenicity in *Erwinia stewartii* require induction by an *N*-acylhomoserine lactone autoinducer. *J. Bacteriol.* **177,** 5000–5008 (1995).
7. Costa, J. M. & Loper, J. E. EcbI and EcbR: homologs of LuxI and LuxR affecting antibiotic and exoenzyme production by *Erwinia carotovora* subsp. *betavasculorum*. *Can. J. Microbiol.* **43,** 1164–1171 (1997).
8. Dunphy, G., Miyamoto, C. & Meighen, E. A homoserine lactone autoinducer regulates virulence of an insect-pathogenic bacterium, *Xenorhabdus nematophilus* (*Enterobacteriaceae*). *J. Bacteriol.* **179,** 5288–5291 (1997).
9. Lewenza, S., Conway, B., Greenberg, E. P. & Sokol, P. A. Quorum Sensing in *Burkholderia cepacia*: identification of the LuxRI Homologs CepRI. *J. Bacteriol.* **181,** 748–756 (1999).
10. Fray, R. G. *et al.* Plants genetically modified to produce *N*-acylhomoserine lactones communicate with bacteria. *Nature Biotech.* **17,** 1017–1020 (1999).
11. Dong, Y.-H., Xu, J.-L., Li, X.-C. & Zhang, L. H. AiiA, a novel enzyme inactivates acyl homoserine-lactone quorum-sensing signal and attenuates the virulence of *Erwinia carotovora*. *Proc. Natl Acad. Sci. USA* **97,** 3526–3531 (2000).
12. von Bodman, S. B., Majerczak, D. R. & Coplin, D. L. A negative regulator mediates quorum-sensing control of exopolysaccharide production in *Pantoea stewartii* subsp. *stewartii*. *Proc. Natl Acad. Sci. USA* **95,** 7687–7692 (1998).
13. Welch, M. *et al.* *N*-acyl homoserine lactone binding to the CarR receptor determines quorum-sensing specificity in *Erwinia*. *EMBO J.* **19,** 631–641 (2000).
14. Qin, Y. P. *et al.* Quorum-sensing signal binding results in dimerization of TraR and its release from membranes into the cytoplasm. *EMBO J.* **19,** 5212–5221 (2000).
15. Pearson, J. P., Van Delden, C. & Iglewski, B. H. Active efflux and diffusion are involved in transport of *Pseudomonas aeruginosa* cell-to-cell signals. *J. Bacteriol.* **181,** 1203–1210 (1999).
16. Zhu, J. & Winans, S. C. Autoinducer binding by the quorum-sensing regulator TraR increases affinity for target promoters *in vitro* and decreases TraR turnover rates in the whole cells. *Proc. Natl Acad. Sci. USA* **96,** 4832–4837 (1999).
17. Leadbetter, J. R. & Greenberg, E. P. Metabolism of acyl-homoserine lactone quorum-sensing signals by *Variovorax paradoxus*. *J. Bacteriol.* **182,** 6921–6926 (2000).
18. Pearson, J. P. *et al.* Structure of the autoinducer required for expression of *Pseudomonas aeruginosa* virulence genes. *Proc. Natl Acad. Sci. USA* **91,** 197–201 (1994).
19. Winson, M. K. *et al.* Multiple *N*-acyl-L-homoserine lactones signal molecules regulate production of virulence determinants and secondary metabolites in *Pseudomonas aeruginosa*. *Proc. Natl Acad. Sci. USA* **92,** 9427–9431 (1995).
20. Zhang, L. H., Murphy, P. J., Kerr, A. & Tate, M. E. *Agrobacterium* conjugation and gene regulation by *N*-acyl-L-homoserine lactones. *Nature* **362,** 446–447 (1993).
21. Gallois, P. & Marinho, P. in *Plant Gene Transfer and Expression Protocols* (ed. Jones, H.) 39–48 (Humana, Totowa, New Jersey, 1995).
22. Beaujean, A., Sangwan, R. S., Lecardonnel, A. & Sangwan-Norreel, B. S. *Agrobacterium*-mediated transformation of three economically important potato cultivars using sliced internodal explants: an efficient protocol of transformation. *J. Exp. Bot.* **49,** 1589–1595 (1998).
23. Borisjuk, N. V. *et al.* Calreticulin expression in plant cells: developmental regulation, tissue specificity and intracellular distribution. *Planta* **206,** 504–514 (1998).
24. Givskov, M. *et al.* Eukaryotic interference with homoserine lactone-mediated prokaryotic signaling. *J. Bacteriol.* **178,** 6618–6622 (1996).
25. Manefield, M. *et al.* Evidence that halogenated furanones from *Delisea pulchra* inhibit acylated homoserine lactone (AHL)-mediated gene expression by displacing the AHL signal from its receptor protein. *Microbiol.* **145,** 283–291 (1999).

© 2001 Macmillan Magazines Ltd

26. Manefield, M., Harris, L., Rice, S. A., de Nys, R. & Kjelleberg, S. Inhibition of luminescence and virulence in the black tiger prawn (*Penaeus monodon*) pathogen *Vibrio harveyi* by intercellular signal antagonists. *Appl. Environ. Microbiol.* **66,** 2079–2084 (2000).
27. Hoang, T. T. & Schweizer, H. P. Characterization of *Pseudomonas aeruginosa* enoyl-acyl carrier protein reductase (FabI): a target for the antimicrobial triclosan and its role in acylated homoserine lactone synthesis. *J. Bacteriol.* **181,** 5489–5497 (1999).
28. Borisjuk, N. V. *et al.* Production of recombinant proteins in plant root exudates. *Nature Biotechnol.* **17,** 466–469 (1999).

Acknowledgements

We thank A. Kerr, M. E. Tate, V. Sundaresan and K. Sampath for critical reviews of the manuscript. Funding was provided by the National Science and Technology Board of Singapore.

Correspondence and requests for materials should be addressed to L.-H.Z. (e-mail: lianhui@ima.org.sg). The 16S rDNA sequence of *Bacillus* sp. 240B1 has been deposited in GenBank under accession number AF350926.

Production of multiple plant hormones from a single polyprotein precursor

Gregory Pearce, Daniel S. Moura, Johannes Stratmann* & Clarence A. Ryan

Institute of Biological Chemistry, Washington State University, Pullman, Washington 99164-6340, USA

Some animal and yeast hormone genes produce prohormone polypeptides that are proteolytically processed to produce multiple copies of hormones with the same or different functions[1]. In plants, four polypeptides have been identified that can be classed as hormones[2–5] (intercellular chemical messengers[6]) but none are known to be produced as multiple copies from a single precursor. Here we describe a polyprotein hormone precursor, present in tobacco plants, that gives rise to two polypeptide hormones, as often found in animals and yeast. The tobacco polypeptides activate the synthesis of defensive proteinase-inhibitor proteins in a manner similar to that of systemin, an 18-amino-acid polypeptide found in tomato plants[2]. The two tobacco polypeptides are derived from each end of a 165-amino-acid precursor that bears no homology to tomato prosystemin. The data show that structurally diverse polypeptide hormones in different plant species can serve similar signalling roles, a condition not found in animals or yeast.

Following the discovery of systemin in tomato leaves, homologues were identified in potato, nightshade and pepper[7], but not in tobacco. Tobacco exhibits a systemic wound induction of proteinase inhibitors[8] similar to that found in tomato, potato and pepper, but we could not detect a systemin-like polypeptide or its precursor in tobacco leaves. (The term 'systemin' is used to describe polypeptide defence signals that are produced by the plant in response to physical damage and that activate defence genes, either locally or systemically.) Systemic responses to herbivore attacks have been documented in more than 100 species of plants[9], and the chemical nature of these signals is pertinent to the evolution of information-processing systems in plants as well as agricultural, biotechnological, environmental and ecological applications. We therefore initiated a search for the tobacco wound signal, hoping that the isolation of a tobacco systemin would provide a clue to the diversity of systemic wound signals within the Solanaceae, and would lead to the identification of polypeptide defence signals in other plant families.

We developed an assay to identify and isolate the systemin-like polypeptide of tobacco with suspension-cultured tobacco cells, on the basis of the ability of tomato systemin to cause an increase in the pH (alkalinization) of the culture medium[10,11]. Aliquots of cultured cells (1 ml) were used to assay alkalinization activity in either crude extracts or in individual fractions from chromatographic columns during purification (see Supplementary Information 1). Aliquots of 1–10 μl from the fractions were added to the agitated cells and the pH was monitored. The pH of the suspension cell medium increases in response to increasing quantities of a crude tobacco leaf extract (see Supplementary Information 2). Increases in the medium pH in response to aliquots of fractions from high-pressure liquid chromatography (HPLC) (see Supplementary Information 1, Step 4) clearly identified two major peaks and one minor peak that eluted from the column (Fig. 1a). Each major peak was further analysed by HPLC, producing purified polypeptides I and II (Fig. 1b; see Supplementary Information 1, Step 6).

The biological activities of polypeptides I and II were assayed in tobacco suspension cultures and in leaves of tobacco plants. The alkalinization of the tobacco cell cultures in response to both polypeptides peaked at about 15 min (Fig. 2a) and then slowly declined. Tobacco polypeptides I and II exhibited similar concentration dependence for the induction of alkalinization of cell culture medium (Fig. 2b) and for the induction of tobacco trypsin inhibitors in leaves (Fig. 2c). Tobacco trypsin inhibitors are members of the potato inhibitor II family[8,12]. Members of this family are induced by wounds, methyl jasmonate and systemin in tomato[13]. Tomato systemin did not cause an alkalinization of the tobacco cell-culture medium, nor did it activate the synthesis of tobacco trypsin inhibitor in leaves when supplied to young tobacco plants through their cut stems, confirming that tomato systemin is not perceived by tobacco cells. Whether tobacco systemins are local signals, systemic signals or both has not yet been established. In tomato plants, an antisense prosystemin severely reduced systemic signalling[14]. Transformation of tobacco plants with an antisense precursor gene for tobacco systemin should be helpful in this regard.

Polypeptides I and II contained 18 amino acids each and were

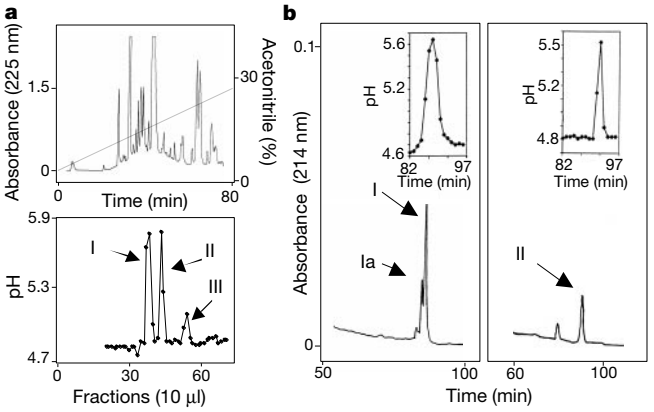


Figure 1 Isolation of biologically active tobacco polypeptides from tobacco leaf extracts. **a**, Reversed phase C18 HPLC separations of partially purified tobacco leaf extracts (top panel) and the assay of 10 μl of each eluted fraction (2 ml) for their ability to cause the alkalinization of 1 ml of the medium of tobacco suspension cultured cells (lower panel). The pH of the culture medium was recorded after 15 min. **b**, Final purification of polypeptides in peak I (left) and peak II (right) from **a** by narrow-bore reversed-phase chromatography. From each 0.25-ml fraction, 1 μl was assayed for its alkalinization activity (insets).

* Present address: Department of Biological Sciences, University of South Carolina, Columbia, South Carolina 29208, USA.

© 2001 Macmillan Magazines Ltd

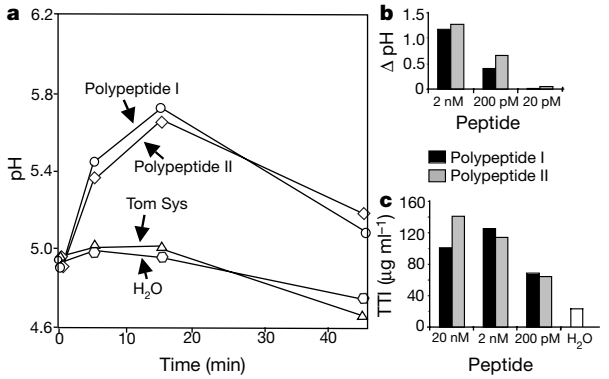


Figure 2 The biological activities of polypeptides I and II. **a**, Time course of alkalinization of the medium of tobacco cell suspension cultures by 1 nM purified tobacco polypeptides from peaks I and II in Fig. 1b, and 1 μM tomato systemin (Tom Sys). **b**, The concentration dependence of the activities of peaks I and II when assayed for alkalinization of the medium of 1 ml of tobacco cell suspension cultures. **c**, The concentration dependence of the proteinase-inhibitor-inducing activities of peaks I and II when supplied to young tobacco plants through their cut petioles for 30 min in 100 μl water, followed by incubation under 200 μE $M^{-2}s^{-1}$ of light for 24 h. Levels of tobacco trypsin inhibitor (TTI) in leaves were quantified by radial immunodiffusion assays with rabbit anti-tobacco trypsin inhibitor serum[9].

named Tobacco Systemin I (Tob Sys I) and Tobacco Systemin II (Tob Sys II) (Table 1). Neither of the tobacco polypeptides is homologous with tomato systemin[2], but -OOS- motifs found in the tobacco polypeptides originate from -PPS- primary translation products that are identical to the -PPS- motifs found in tomato systemin. Mass spectroscopic analyses (see Supplementary Information 3) confirmed that both polypeptides are linked to carbohydrate residues. The tobacco systemins may be related to hydroxyproline-rich glycoproteins that are structural components of plant cell walls[15]. The -PPSPX- motif found in both Tob Sys I and II amino-acid sequences has been identified as a repeating unit in a hydroxyproline-rich glycoprotein, GPI, that may be involved in cell-wall assembly[16]. Mass spectroscopy of products during time courses of acid hydrolyses of Tob Sys I and II polypeptides revealed a ladder of fragments consistent with the carbohydrates being composed entirely of pentose units (see Supplementary Information 3). These analyses indicated that the polypeptides contained nine pentose units (Tob Sys I) and six pentose units (Tob Sys II), respectively (Table 1; see Supplementary Information 3).

The small peak III from Fig. 1a was nearly identical to Tob Sys II, differing only by having an additional Leu at the carboxy terminus. Polypeptide III exhibited activity similar to Tob Sys I and II in the alkalinization assay, indicating that the carboxy-terminal Leu did not affect the activity. Analysis of peak Ia in Fig. 1b yielded an amino-acid sequence nearly identical to that of Tob Sys I, except for a substitution of Thr for Ala at position 3. These differences became clear when the complementary DNAs were isolated, as described below.

In Tob Sys I and II, the carbohydrates were released by mild acid hydrolysis, but the linkages and identities of the carbohydrates have not been established. Removal of the carbohydrates of each tobacco systemin by mild acid hydrolysis resulted in molecular masses identical to the calculated masses determined by sequence analysis (Table 1; see Supplementary Information 3). Deglycosylated polypeptides could not be recovered in quantities for assay, but synthetic Tob Sys I and II were 10,000 times less active than the native polypeptides in the alkalinization assay (Fig. 2b; see Supplementary Information 4), indicating that the carbohydrate moieties are important for activity, but are not absolutely required. The presence of hydroxyproline residues and carbohydrates in the polypeptides indicated that their synthesis involves the secretory pathway.

The physical properties of Tob Sys I and II are in striking contrast to those of tomato systemin, which has no carbohydrate residues, prolines that are not hydroxylated and a precursor with no signal sequence[2,14]. An oligonucleotide primer, (A/G)-C-C-(A/T/C/G)-C-T-(A/T/C/G)-T-C-(A/T)-C-C-A-C-C-A-T-C-(A/T)-C, deduced from the amino-acid sequence of Tob Sys I, -PLSOOS- (hydroxyprolines being post-translational modifications of prolines), was synthesized and used to isolate a precursor cDNA with reverse-transcription polymerase chain reaction. The deduced preproprotein (Fig. 3) coded for a protein of 165 amino acids that included a signal sequence. Within this protein are the sequences of Tob Sys I and Tob Sys II, with Tob Sys I near the amino-terminus and Tob Sys II near the C terminus. This is the first example of a plant polypeptide prohormone that produces multiple polypeptide signals. Standard protein–protein and nucleotide–nucleotide BLAST (Basic Local Alignment Search Tool; http:// www.ncbi.nlm.nih.gov/BLAST) searches for cDNA and preproprotein homologies did not reveal any significant similarities or identities with any genes or proteins, including hydroxyproline-rich glycoproteins. Extensive probing of several tobacco-leaf cDNA libraries revealed the presence of a second tobacco systemin precursor (see Supplementary Information 5) that is 164 amino acids in length and differs from the initially isolated cDNA at 19 amino-acid sites. One of the variable sites was in the sequence of Tob Sys I that confirmed the Thr substitution for Ala found in the amino-acid sequence analyses of peak Ia (Fig. 1b). This difference in the sequences indicates that Tob Sys I in peak I is derived from pro-TobSys-A, whereas peak Ia is derived from pro-TobSys-B (see Supplementary Information 5). The differences of 30 mass units are reflected in the Maldi mass spectral analyses (see Supplementary Information 3). Southern analyses (see Supplementary Information 6) confirmed that the tobacco genome contains a small gene family of the precursor gene that is probably composed of just two members.

Exposing young tobacco plants to methyl jasmonate caused an increase in tobacco systemin precursor messenger RNAs within 6 h (see Supplementary Information 7), which explains the 2.5-fold increase in tobacco systemin levels in leaves treated with methyl jasmonate and subsequently used for purification. This indicates that the gene is regulated through the octadecanoid pathway, similar to the regulation of the prosystemin gene in tomato leaves[13], where the gene is constitutively expressed at a low level (see Supplementary Information 7).

Tob Sys I and II caused a rapid activation of a mitogen-activated protein (MAP) kinase with relative molecular mass 48,000 (M_r 48K), similar to the 48K MAP kinase activated by tomato systemin in tomato cells (Fig. 4)[17]. Tob Sys II activated a MAP kinase earlier and stronger than Tob Sys I, and the kinase remained active for a longer period, indicating a difference between the polypeptides in perception and/or intracellular signalling. However, the differences observed in MAP kinase activation are not reflected in the alka-

Table 1 Amino-acid sequences and relative molecular masses of tobacco systemins

Peptides	Amino-acid sequence	Mass	Mass after hydrolysis	Mass lost	Divide by 132 (pentose mass)
Tob Sys I	$^{+}$RGANLPOOSOASSOOSKE^{-}	3,060	1,869	1,191	9
Tob Sys II	$^{+}$NRKPLSOOSOKPADGQRP^{-}	2,784	1,992	792	6
Tom Sys	$^{+}$AVQSKPPSKRDPPKMQTD^{-}	2,010	ND	ND	ND

Tob Sys, tobacco systemin; Tom Sys, tomato systemin; ND, not determined.

© 2001 Macmillan Magazines Ltd

```
     TCGACGACCAAATTGATTCTCATAATGAGAGTTCTGTTTCTCATCTACCTTATCCTCTCT

  1                          M  R  V  L  F  L  I  Y  L  I  L  S   12-
     CCATTTGGAGCTGAAGCAAGAACTTTGCTAGAAAATCATGAGGGGCTGAACGTAGGATCT

 13  P  F  G  A  E  A  R  T  L  L  E  N  H  E  G  L  N  V  G  S   32-
     GGTTATGGAAGGGGTGCAAATCTACCACCACCATCTCCAGCTTCATCTCCACCGAGCAAG
              TOB SYS I
 33  G  Y  G  R  G  A  N  L  P  P  P  S  P  A  S  S  P  P  S  K   52-
     GAGGTATCCAACAGTGTGTCTCCTACTCGCACTGATGAGAAAACCTCTGAAAACACTGAG

 53  E  V  S  N  S  V  S  P  T  R  T  D  E  K  T  S  E  N  T  E   72-
     CTCGTTATGACAACAATTGCTCAAGGGGAAAACATTAATCAGCTGTTTTCATTTCCAACT

 73  L  V  M  T  T  I  A  Q  G  E  N  I  N  Q  L  F  S  F  P  T   92-
     TCAGCAGACAATTACTACCAATTAGCAAGTTTCAAGAAACTGTTCATCTCATACCTTCTG

 93  S  A  D  N  Y  Y  Q  L  A  S  F  K  K  L  F  I  S  Y  L  L  112-
     CCTGTCTCATATGTATGGAACCTGATCGGTTCATCGTCTTTTGATCACGATCTTGTAGAT

113  P  V  S  Y  V  W  N  L  I  G  S  S  S  F  D  H  D  L  V  D  132-
     ATTTTTGACTCAAAGTCCGATGAAAGATACTGGAATCGTAAGCCTCTATCACCACCATCT
                                          TOB SYS II
133  I  F  D  S  K  S  D  E  R  Y  W  N  R  K  P  L  S  P  P  S  152-
     CCAAAGCCAGCAGATGGTCAGCGTCCTCTTCACTCCTATTAAGCCTCGTATCTCTAGTTT

153  P  K  P  A  D  G  Q  R  P  L  H  S  Y  *                   165-
     GGAGTTTTATTATGTGTAGTACCAGGTACTAAGTTAGATTTTATTTGGTACTTTTGTAAG
     CTGCTTTAAGTTCACTTATCAATTATGGTAATGGTCGAATAACTTTTAGCAATTTAATAA
     TATGGCTATAGGGTTGTAATTTAGCGGGTCGTTTGGTTTTGATATAGTAATAAGTTAAGC
     GTGAATTAATAATATAGATAATATTGTAATGTGTATTTCTTGTTTGATGTTTGGTTTGAT
     GTATTAATAATTACTAATGCGTAATCAAAGATGGTGTGTGTTTTGTTTAAAAAAAAAAAA
```

Figure 3 The nucleotide and amino-acid sequence of the Tob Sys I and II polyprotein precursor. The Tob Sys I and II polypeptide sequences (double underline) are derived from the N-terminal and C-terminal regions of the tobacco preproprotein. The deduced signal sequence is underlined.

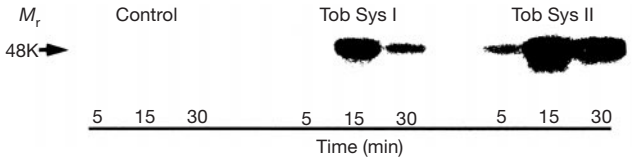


Figure 4 Activation of a MAP kinase with M_r 48K in tobacco suspension-cultured cells by Tob Sys I and II. Activity of MAP kinase was measured as previously described[17,18]. Tob Sys I and II polypeptides (2 nM) were added to 1 ml suspension culture and the cells were collected and homogenized in SDS–PAGE sample buffer; the proteins were separated by SDS–PAGE in which the gel contained myelin basic protein. The gels were treated with ^{32}P-ATP and washed, and the phosphorylated myelin basic protein was identified by radio autography.

linization response or in the accumulation of the tobacco trypsin inhibitor induced by the two polypeptides (Fig. 2).

Our data show that polypeptide hormones in plants can arise from a polyprotein precursor, a pattern often found in animals and yeast. Processing of prohormones in animals and yeast usually involves cleavage of peptide bonds at the C terminus of a dibasic pair, but neither the tobacco prosystemin gene product nor any of the other known polypeptide signals in plants exhibit dibasic processing sites. Although the study of polypeptide signalling in plants is in its infancy, it is clear that polypeptide hormones are involved in several aspects of plant growth, differentiation and defence, and are likely to be involved in many plant processes. Putative receptors for four of the known polypeptide hormones in plants have been identified[18–21], and it will be of interest to know whether Tob Sys I and II are mediated by the same or independent cell surface receptors. The further identification of polypeptide hormones and their receptors in plants will probably continue to reveal analogies with animals and yeast, providing insight into the evolution of polypeptide signalling systems from a common ancestor. □

Methods

Assay

Tobacco suspension cells were maintained in Murashige and Skoog medium as described[22], but excluding buffer. Instead, the medium was adjusted to pH 5.6 with KOH. We transferred 3-ml aliquots of one-week-old cultures into 35 ml of medium in 125-ml flasks and placed them on an orbital shaker at 160 r.p.m. in the dark for growth. Cells were used for assay 3–5 d after transfer (cell density was $\sim 0.5–1.5 \times 10^6$ cells ml^{-1}). A 1-ml aliquot of cells was transferred into each well of 24-well cell-culture cluster plates (Costar, Corning Incorporated) and allowed to equilibrate on an orbital shaker at 160 r.p.m. for 1 h. Aliquots (1–10 μl) of fractions to be tested were added to the cells and the increasing pH change of the medium (medium alkalinization) was detected over time with an Orion EA940 pH meter with an Orion semi-micro pH electrode.

Polypeptide isolation

Tobacco plants were collected for extraction four weeks after planting. The plants were sprayed with a solution of methyl jasmonate (125 μl in 500 ml distilled water containing 0.125% Triton X-100) and then incubated under constant light for 15 h. This treatment caused about a 2.5-fold increase in activity of each polypeptide. Plants were collected in liquid N_2 and stored at –20 °C until use. Each preparation consisted of approximately 450 plants with a wet weight of about 1.1 kg. The leaves were homogenized in a 4-l Waring blender with 2.8 l of 1% trifluoroacetic acid for 2 min and filtered through eight layers of cheesecloth and one layer of Miracloth (Calbiochem). The liquid was centrifuged at 10,000*g* for 20 min.

Polypeptides in the supernatant from the initial centrifugation were separated with C18 reversed phase chromatography (for full details, see Supplementary Information 1). The 30% methanol-eluting fraction, called 'crude polypeptide extract' was found to contain tobacco inhibitor-inducing activity, and it also had strong activity in alkalinizing the media of tobacco suspension cells (see Supplementary Information 2). The methanol was removed with a rotary evaporator, followed by lyophilization to dryness. The average yield per preparation was about 390 mg. Several preparations were collected and stored to provide a stock for further purification. Active fractions were further separated on Sephadex G-25. The active fractions were located at or close to the void volume. This fraction was pooled and lyophilized. The average yield per preparation was 46 mg. The lyophilized powder (25 mg) was further separated by semi-preparative reversed-phase C18 chromatography. Activity peaks I, II and III were lyophilized (Fig. 1a). Strong cation-exchange chromatography was performed on the active peaks from the C18 semi-

© 2001 Macmillan Magazines Ltd

preparative column. The active fractions were pooled and lyophilized. Narrow-bore reversed-phase chromatography was performed with each of the active peaks (Fig. 1b). The peak areas were compared with the area of 250 pmol of tomato systemin standard and a 13-amino-acid synthetic tobacco polypeptide standard. The yield of peak Ia was 230 pmol and of peak I was 752 pmol. Peak II yielded 124 pmol. This was the total yield from 50 mg of crude G-25 purified leaf extract.

N-terminal sequencing was performed by Edman chemistry on an Applied Biosystems Model 475 sequencer with pulse liquid update using the manufacturer's protocol. C-terminal sequence analyses were performed on polypeptides digested with carboxypeptidase P and analysed by MALDI mass spectrometry[23]. Samples were prepared for carboxypeptidase P digestion by first treating samples with 1% trifluoroacetic acid at 80 °C for 3 h to hydrolyse the pentose residues from the peptides. The C terminus of peak I was -(K/Q)E, and the N terminus was RGANLPOOSOASSOOSKE. The C terminus of peak II was -DG(K/Q)RP, and the N terminus was NRKPLSOOSOKPADGQ-. A partial N-terminal sequence of peak III determined that the first five N-terminal residues matched the N terminus of the peak II sequence, and the mass was 113 units larger than Tob Sys II. This indicated that a leucine was probably present at its C terminus, as expected from the deduced amino-acid sequence of the preproteins (Fig. 3).

Received 5 December 2000; accepted 30 April 2001.

1. Niall, H. D. The evolution of peptide hormones. *Ann. Rev. Physiol.* **44,** 615–624 (1982).
2. Pearce, G., Strydom, D., Johnson, S. & Ryan, C. A. A polypeptide from tomato leaves induces wound-inducible proteinase inhibitor proteins. *Science* **253,** 895–897 (1991).
3. Matsubayashi, Y. & Sakagami, Y. Phytosulfokine, sulfated peptides that induce the proliferation of single mesophyll cells of *Asparagus officinalis* L. *Proc. Natl Acad. Sci. USA* **93,** 7623–7627 (1996).
4. Fletcher, J. C., Brandu, U., Running, M. P., Simon, R. & Meyerowitz, E. M. Signaling of cell fate decisions by CLAVATA3 in *Arabidopsis* shoot meristems. *Science* **283,** 1911–1914 (1999).
5. Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. The male determinant of self-incompatibility in *Brassica*. *Science* **286,** 1697–1700 (1999).
6. Voet, D. & Voet, J. G. *Biochemistry* 2nd edn 1261 (Wiley and Sons, New York, 1995).
7. Constabel, C. P., Yip, L. & Ryan, C. A. Prosystemin from potato, black nightshade, and bell pepper: primary structures and biological activities of the predicted systemins. *Plant Mol. Biol.* **26,** 55–62 (1998).
8. Pearce, G., Johnson, S. & Ryan, C. A. Purification and characterization from tobacco (*Nicotiana tabacum*) leaves of six small, wound-inducible, proteinase isoinhibitors of the potato inhibitor II family. *Plant Physiol.* **102,** 639–644 (1993).
9. Karban, R. & Baldwin, I. T. *Induced Responses to Herbivory* (Univ. Chicago Press, Chicago, 1997).
10. Meindl, T., Boller, T. & Felix, G. The plant wound hormone systemin binds with the N-terminal part to its receptor but needs the C-terminal part to activate it. *Plant Cell* **10,** 1561–1570 (1998).
11. Schaller, A. & Oecking, C. Modulation of plasma membrane H^+-ATPase activity differentially activates wound and pathogen defense responses in tomato plants. *Plant Cell* **11,** 263–272 (1999).
12. Bryant, J., Green, T., Gurusaddaiah, T. & Ryan, C. A. Proteinase inhibitor II from potatoes: isolation and characterization of its protomer components. *Biochemistry* **15,** 3418–3424 (1976).
13. Ryan, C. A. The systemin signaling pathway: differential activation of plant defensive genes. *Biochim. Biophys. Acta* **1477,** 112–121 (2000).
14. McGurl, B., Pearce, G., Orozco-Cardenas, M. & Ryan, C. A. Structure, expression and antisense inhibition of the systemin precursor gene. *Science* **255,** 1570–1573 (1992).
15. Sommer-Knudsen, J., Bacic, A. & Clarke, A. E. Hydroxyproline-rich plant glycoproteins. *Phytochemistry* **47,** 483–497 (1998).
16. Ferriss, P. J. *et al.* Glycosylated polyproline II rods with kinks as a structural motif in plant hydroxyproline-rich glycoproteins. *Biochemistry* **40,** 2978–2987 (2001).
17. Stratmann, J. W. & Ryan, C. A. Myelin basic protein kinase activity in tomato leaves is induced systemically by wounding and increases in response to systemin and oligosaccharide elicitors. *Proc. Natl Acad. Sci. USA* **94,** 11085–11089 (1997).
18. Scheer, J. M. & Ryan, C. A. A 160 kDa systemin receptor on the surface of *Lycopersicon peruvanium* suspension-cultured cells. *Plant Cell* **11,** 1525–1535 (1999).
19. Matsubayashi, Y. & Sakagami, Y. 120- and 160-kDa receptors for endogenous mitogenic peptide, phyosulfokine-α, in rice plasma membranes. *J. Biol. Chem.* **275,** 15520–15525 (2000).
20. Trotochaud, A. E., Hao, T., Wu, G., Yang, Z. & Clark, S. E. The CLAVATA1 receptor-like kinase requires CLAVATA3 for its assembly into a signaling complex that includes KAPP and a Rho-related protein. *Plant Cell* **11,** 393–405 (1999).
21. Stein, J. C., Howlett, B., Boyes, D. C., Nasrallah, M. E. & Nasrallah, J. B. Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleraceae*. *Proc. Natl Acad. Sci. USA* **88,** 8816–8820 (1991).
22. Dombrowski, J. E., Gomez, L., Chrispeels, M. & Raikel, N. V. in *Plant Molecular Biology Manual* (eds Gelvin, S. B. & Schilperoort, R. A.) J3, 1–29 (Kluwer Academic, Dordrecht, 1994).
23. Patterson, D. H., Tarr, G. E., Regnier, F. E. & Martin, S. A. C-terminal ladder sequencing via matrix-assisted laser desorption mass spectrometry coupled with carboxypeptidase Y time-dependent and concentration-dependent digestions. *Anal. Chem.* **67,** 3971–3978 (1995).

Supplementary information is available on *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

This research was supported by the College of Agriculture and Home Economics and by the National Science Foundation. We thank S. Vogtman for growing plants; G. Munske for amino-acid sequence analyses; and W. Siems for MALDI mass spectroscopic analyses.

Correspondence and requests for materials should be addressed to C.A.R. (e-mail: cabudryan@hotmail.com). The GenBank accession number for tobacco systemin precursor pro-TobSys-A is AY033148, and for pro-TobSys-B is AY033149.

Structure of a human γδ T-cell antigen receptor

Timothy J. Allison*, Christine C. Winter*, Jean-Jacques Fournié†, Marc Bonneville‡ & David N. Garboczi*

* *Structural Biology Section, Laboratory of Immunogenetics, National Institute of Allergy and Infectious Diseases, 12441 Parklawn Drive, Rockville, Maryland 20852, USA*
† *INSERM U395, CHU Purpan BP3028, 31024 Toulouse, France*
‡ *INSERM U463, Institut de Biologie, 44035 Nantes, France*

T-cell antigen receptors composed of γ and δ polypeptide chains (γδTCRs) can directly recognize antigens in the form of intact proteins or non-peptide compounds, unlike αβTCRs, which recognize antigens bound to major histocompatibility complex molecules (MHC). About 5% of peripheral blood T cells bear γδTCRs, most of which recognize non-peptide phosphorylated antigens[1,2]. Here we describe the 3.1 Å resolution structure of a human γδTCR from a T-cell clone[3] that is phosphoantigen-reactive. The orientation of the variable (V) and constant (C) regions of the γδTCR is unique when compared with αβTCRs or antibodies, and results from an unusually small angle between the Vγ and Cγ domains. The complementarity-determining regions (CDRs) of the V domains exhibit a chemically reasonable binding site for phosphorylated antigens, providing a possible explanation for the canonical usage of the *Vγ9* and *Vδ2* gene segments by phosphoantigen-reactive receptors. Although the γδTCR V domains are similar in overall structure to those of αβTCRs, γδTCR C domains are markedly different. Structural differences in Cγ and Cδ, and in the location of the disulphide bond between them, may enable γδTCRs to form different recognition/signalling complexes than αβTCRs.

γδTCRs, αβTCRs and antibodies are the three types of immune system receptors that are products of rearranging gene segments. Antigen recognition by γδTCRs resembles recognition by antibodies[1,4]. γδ T cells recognize small molecules and intact proteins without the requirement for antigen processing that αβ T cells exhibit[2,5]. Small-molecule recognition by γδ T cells requires cell–cell contact[6,7], suggesting that non-MHC molecules may present small antigens to γδTCRs or that co-stimulation from neighbouring cells is required. Most γδ T cells bear a tissue-specific, restricted set of *Vγ* and Vδ domains. For example, 50–90% of human γδ T cells derived from peripheral blood of healthy donors use a combination of *Vγ9* and *Vδ2* gene segments[8]. These γδTCRs are polyclonal with enormous overall sequence diversity at the junctions of their variable, diversity (*D*) and joining (*J*) segments, but with amino-acid similarities at a few junctional positions. *In vitro*, all $V\gamma9V\delta2^+$ T cells bearing these sequence similarities can be activated by soluble extracts of *Mycobacteria*, *Plasmodia* and other pathogens[9]. Like most $V\gamma9V\delta2^+$ T cells derived from blood, G115 T cells proliferate, secrete lymphokines and lyse target cells in response to stimulation with various natural and synthetic phosphoantigens[3,7,8]. The purified, highly active antigens from *Mycobacteria* are small, phosphate-containing compounds, such as 3-formyl-1-butyl-pyrophosphate[3,10]. $V\gamma9V\delta2^+$ T cells are also stimulated by a variety of less active compounds including isopentenyl pyrophosphate[11], alkylamines[12] and aminobisphosphonates[13].

Like αβTCRs and antigen-binding fragments of antibodies (Fabs), the G115 γδTCR is made up of two polypeptide chains each having two immunoglobulin-like domains: an amino-terminal V domain and a carboxy-terminal C domain (Fig. 1a). The Vγ and Vδ domains are each composed of a 'sandwich' of two β-sheets with five inner and four outer strands. The γ-chain and δ-chain CDR

© 2001 Macmillan Magazines Ltd

loops at the top of the V domains project out from the receptor (Fig. 1b). The Cγ and Cδ domains are made up of two β-sheets with four inner and three outer strands. The four molecules in the asymmetric unit of the G115 crystal are identical to each other except for small displacements of loops that make contact with neighbouring molecules to form the crystal; the angles between the domains of each receptor are essentially the same.

The G115 γδTCR has a distinctive overall shape compared with αβTCRs and Fabs, in that the C domains of the γδTCR are swung out from under the V domains (Fig. 1). This unusual shape is highlighted by both a small elbow angle and a small Vγ–Cγ inter-domain angle. In antibodies and TCRs, the 'elbow angle' is defined as the angle between the pseudo two-fold symmetry axes that relate V to V and C to C. The elbow angle for G115 is 110° and is smaller than any elbow angle observed in αβTCRs or Fabs (Fig. 2). Analysis of 251 Fab and 11 αβTCR molecules revealed average elbow angles of 159° ($\sigma = 20°$) and 149° ($\sigma = 6°$), respectively, with a range from 125° (Protein Data Bank code 1sbs) to 225° (8fab) for the Fabs and from 140° (1nfd) to 159° (1tcr) for the αβTCRs.

The Vγ–Cγ inter-domain angle of 42°, between the long axis of the Cγ domain and the long axis of the Vγ domain, is markedly smaller than the average angle between the V and C domains of Fabs or αβTCRs (Fig. 1b, γ-chain in gold). The Vδ–Cδ inter-domain angle is 101°. Antibodies and αβTCRs have average angles of 92° (V_L–C_L, $\sigma = 9°$), 76° (V_H–C_H, $\sigma = 11°$), 100° (Vα–Cα, $\sigma = 4°$) and 67° (Vβ–Cβ, $\sigma = 3°$). The smallest angles observed for these molecules are 51°, 50°, 93° and 62° for the light (L), heavy (H), α- and β-chains, respectively. Therefore, the 42° Vγ–Cγ angle of G115 is the smallest inter-domain angle observed among these receptors.

As the relative orientations between Vγ and Vδ and between Cγ and Cδ domains are similar to the relative orientations betweenV–V and between C–C of Fabs and αβTCRs, the small angle between the Vγ and Cγ domains shifts both Cγ and Cδ to one side (Fig. 1b). The G115 γδTCR exhibits 1,077 Å^2 of buried surface in the interface between its V and C domains, similar to that for Fabs and αβTCRs, which average about 900 Å^2 and 1,200 Å^2, respectively. Although the correlation between the amount of buried surface in a V–C interface and the elbow angle is not known, it is expected that the residues in this interface affect the elbow angle. As the residues found in the G115 interface are conserved in most human and mouse γ-chain and δ-chain genes, it is likely that other γδTCRs will have similar elbow angles and a comparable molecular shape.

Sequence similarities between Vγ and Vβ and between Vδ and Vα domains clearly indicate their relatedness, but whether Vγ or Vδ is more like an antibody V_L or V_H domain is more difficult to determine. Both possible superpositions of Vγ and Vδ with V_L or V_H domains yield similar root-mean-square (r.m.s.) deviations in α-carbon positions and do not distinguish which V domain is more like a V_L or a V_H domain. On the basis of elbow angle considerations, we chose to structurally align Vγ with V_H and Vδ with V_L. This yields an elbow angle of 110° for G115 and an angle that is less than 180° for most Fabs and for all structurally determined TCRs (Fig. 2). Alignment in this way of the γδTCR domains with their analogous domains in αβTCRs and in Fabs shows the overall similarity of these immunoglobulin-like domains and highlights their differences (see Fig. 1 of Supplementary Information).

The CDR loops of γδTCRs, αβTCRs and antibodies comprise the sites of antigen recognition. The loop residues of G115 are: γ25–36, δ25–35 in CDR1 loops; γ53–59, δ51–55 in CDR2 loops; γ73–79, δ69–74 in hypervariable-4 loops (HV4s); and γ99–111, δ94–108 in CDR3 loops. The CDR1, CDR2 and HV4 loops of Vγ are similar in length and in relative position to those of Vβ. The CDR1 and HV4 loops of Vδ are also similar to those of Vα. However, the CDR2 loop of Vδ differs in relative position with that of Vα, in which the C″ strand pairs with the D strand of the outer β-sheet. In Vδ, as in Vγ, Vβ, V_L and V_H, the C″ strand is paired with the C′ strand of the inner β-sheet of the domain. This C″–C′ pairing in Vδ has been seen in the structure of an isolated human Vδ3 domain[14].

As expected, the Vδ2 domain of the G115 structure is very similar to the isolated Vδ3 domain[14]; the r.m.s. difference between the two domains is 1.2 Å over 86 α-carbon atoms. There are two significant differences between Vδ2 and Vδ3 domains. The CDR2 (C′–C″) loop of Vδ2 (and Vδ1) is three residues shorter than the equivalent loop in Vδ3. Unlike the solvent-exposed CDR1 loop of Vδ3, the CDR1 loop of Vδ2 contains an isoleucine that tucks into the core of the domain and holds the loop away from CDR3. As the CDR1 loop

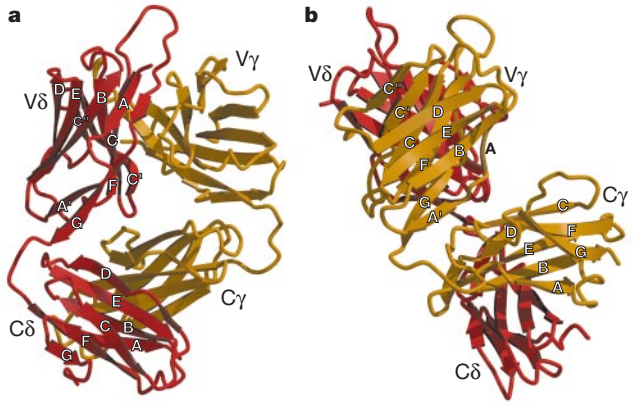


Figure 1 Views of the G115 Vγ9Vδ2 TCR. **a**, View of Vδ and Cδ (red) and Vγ and Cγ (gold) domains. CDR loops are at the top. **b**, View after a 90° rotation around the vertical axis. The nine β-strands of the V domains are labelled A–G, including A′, C′, and C″. The seven β-strands of the C domains are labelled A–G. In each domain, strands A, B, E, and D form one β-sheet, and strands G, F and C (including A′, C′ and C″ in the V domains) form the other sheet. The Vγ–Vδ interface involves the A′GFCC′C″ β-sheets. The Cγ–Cδ interface involves the ABED β-sheets.

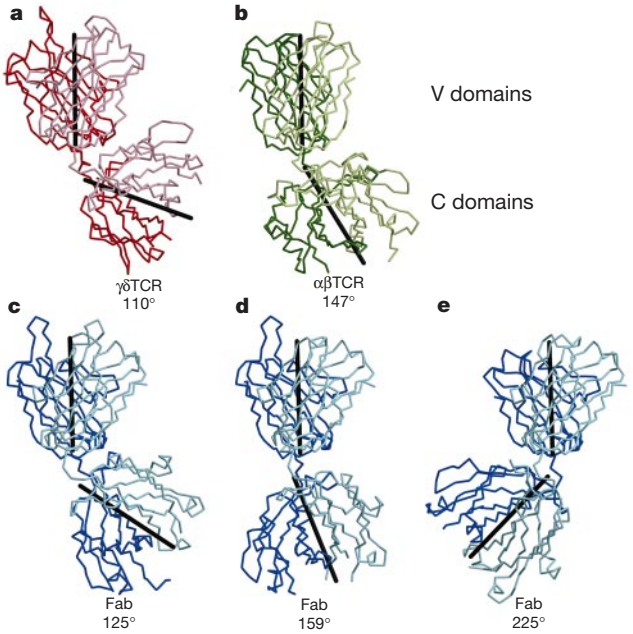


Figure 2 Overall structure of a γδTCR, αβTCR and Fabs. **a**, γδTCR G115 (red); **b**, αβTCR (green); and **c–e**, three Fabs (blue) aligned using both V domains. Note the atypical position of the γδ C domains. The pseudo two-fold symmetry axes (dyads) within the V and C domains are shown (solid lines). The elbow angle between these axes is listed for each representative molecule (αβTCR: Protein Data Bank code 1qsf; Fabs: **c**, 1sbs; **d**, 1hil; **e**, 8fab). The γ-, β- and H chains and the δ-, α- and L chains are in lighter and darker shades, respectively.

© 2001 Macmillan Magazines Ltd

of Vδ1 contains a tryptophan residue at this position, its loop may more closely resemble that of Vδ2 than Vδ3.

The CDR-containing surface of the G115 V domains exhibits more protrusions and clefts compared with the flatter surfaces of αβTCRs that bind a peptide–MHC complex, and of antibodies that bind protein antigens (Fig. 3). The jagged surface of this γδTCR resembles the surface of an antibody that binds a small-molecule antigen. If phosphoantigens are recognized while bound to other molecules, the shape of presentation molecules that would complement this rather rough γδTCR surface would probably be different from the shape of MHC molecules.

Vγ9-bearing T cells respond to the superantigen staphylococcal enterotoxin A (SEA)[15]. Superantigens bind to αβTCRs through the side and top of the Vβ domain; in particular the CDR2β and HV4β loops and nearby framework regions[16]. The structural similarity of Vγ9 and Vβ and the superposition of Vγ9 and SEA with Vβ and staphylococcal enterotoxin B (SEB) of the Vβ–SEB complex suggests that the interaction of Vγ with SEA will resemble that of Vβ with SEB[16], in agreement with a recent report[17].

From the sequences of many TCRs, it has been noted that CDR3γ loops are similar in length to CDR3β loops, and that CDR3δ loops are longer than CDR3α loops[4]. For G115, however, both of its CDR3 loops are similar in length to those of αβTCRs for which the structures have been determined. Unlike αβTCRs, the G115 CDR3 loops protrude significantly from the rest of the molecule and create a cleft between them. Portions of CDR1δ, CDR1γ and CDR2γ combine with the cleft between the CDR3 loops to form a pocket in G115 that is likely to be the phosphoantigen-binding site. In the packing of the crystal, the CDR3δ loop of each molecule makes a contact with Arg γ59 of a crystallographically related molecule that may partially block the cleft and may be responsible for preventing phosphoantigen binding in our crystal-soaking and co-crystallization experiments.

Phosphoantigen molecules are composed of pyrophosphate, a short alkyl carbon chain, and often a chemically reactive group. G115 γδ T cells respond to the *Mycobacterium tuberculosis* antigen 3-formyl-1-butyl-pyrophosphate, which has a pyrophosphate, an alkyl chain and a formyl group that are important for T-cell activation (Fig. 3b)[11,18]. In cellular assays, excess pyrophosphate inhibits activation by phosphoantigens[18]. The possible phosphoantigen-binding site on G115 is the positively charged pocket formed by Arg 59 of CDR2γ, Lys 109 of CDR3γ, and Arg 51 of CDR2δ, which may interact with the phosphates of the antigen (Fig. 3c, d). Similar positively charged pockets are found in antibodies that bind phosphate-containing antigens.

A key residue of the putative binding site is Lys 109 of the CDR3γ loop, which derives from the *JγP* gene segment. In the approximately 50 reported sequences of receptors from γδ T cells activated by phosphoantigens, all use the *Vγ9* gene segment and almost all use the *JγP* gene segment[19,20]. A few reactive clones use the *Jγ1* or *Jγ2* gene segments that encode a lysine residue in a similar position in the CDR3 sequence that may serve the same role as Lys γ109 of JγP. The other *Jγ* gene segments that have not been found in reactive clones, *JγP1* and *JγP2*, do not encode an analogous lysine residue. Of the Vγ-encoded residues that are located near the putative binding pocket, all but two (Thr γ33, Trp γ100) of them are found only in the *Vγ9* gene segment (Fig. 3d); this may explain the absolute requirement for Vγ9 in phosphoantigen-reactive clones.

The CDR3 loops link the ends of the F and G *β*-strands in all V domains. In Vγ9 of G115, the two-stranded FG *β*-sheet extends two residues farther than known F and G strands in Vβ domains. Only four residues at the end of the 13-residue CDR3γ loop are not in regular *β*-sheet structure. About one-third of the phosphoantigen-reactive sequences have no N-encoded residues at the Vγ9–JγP joint, and the others have from one to three N-encoded residues, with G115 having two[8,19].

CDR3δ sequences in γδTCRs are, in general, very diverse as up to three *D* gene segments with N-addition can be used[1]. CDR3δ chains from phosphoantigen-reactive γδTCRs usually bear a single *Dδ3* gene segment but can have from three to twelve residues of diverse sequence encoded by *D*- and N-derived nucleotides[8]. The CDR3δ loop of G115 consists of fifteen residues, seven from *D* and N additions, three from *Vδ2* and five from *Jδ1*. All phosphoantigen-reactive clones have a conserved hydrophobic residue, encoded by N-addition nucleotides, at position 97 of CDR3δ[19] with Leu, Ile, Val and Gly being predominant; G115 has a Leu at δ97 (Fig. 3d). This conserved hydrophobe may interact with hydrophobic parts of

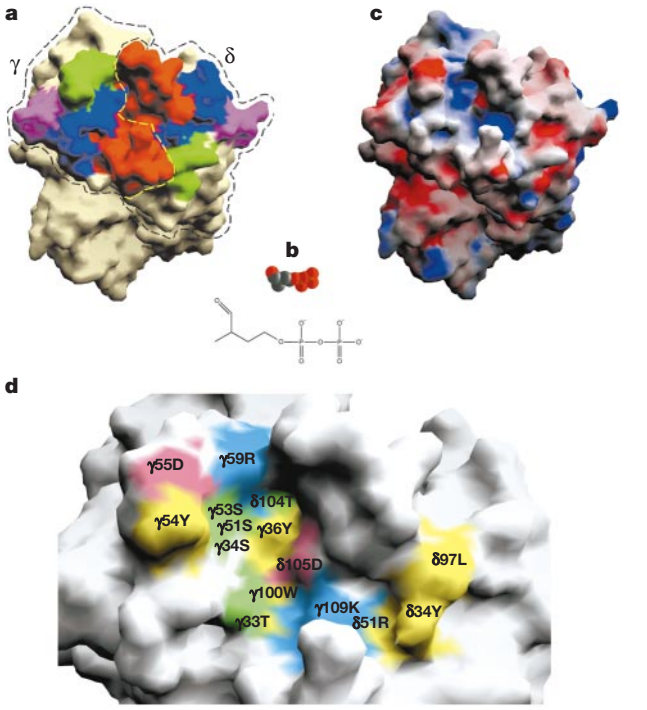


Figure 3 Surface representations of G115. **a**, The surface of G115 is coloured by CDR loops (CDR1, blue; CDR2, green; HV4, pink; and CDR3, red). **b**, 3-formyl-1-butyl-pyrophosphate is shown as a drawing and as a space-filling model (at the same scale as G115 in **a** and **c**). **c**, The surface potential of G115 is colour-coded from red (–10 kT) to blue (+10 kT). The top of the molecule, containing the CDR loops, is facing the viewer. The cleft between the two CDR3 loops contains positively charged patches (blue), which line the phosphoantigen-binding site. **d**, Enlarged view of the cleft shows the location of charged (negative in red; positive in blue), hydrophobic (yellow) and polar (green) residues.

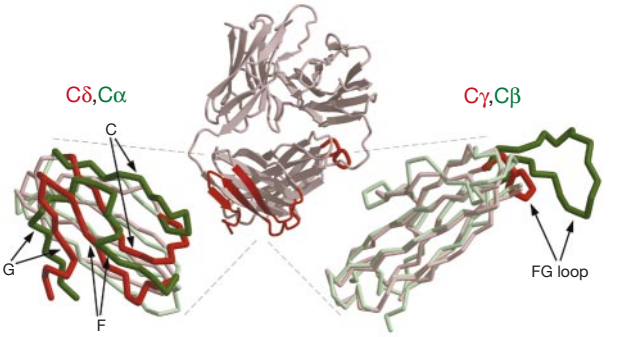


Figure 4 Comparison of the γδTCR and αβTCR C domains. Superpositions of Cδ with Cα (left) and Cγ with Cβ (right) are shown. Differences between the domains are highlighted; the locations of these differences within the G115 structure are also shown. Cδ and Cγ are shown in red. Cα and Cβ from 1qsf are shown in green. The outer face of Cδ contains *β*-strands C, F and G, which form a regular *β*-sheet. Cγ has a much shorter FG loop.

© 2001 Macmillan Magazines Ltd

phosphoantigen molecules and seems to be antigen-selected, as thymic Vγ9Vδ2 TCR sequences are not restricted to being hydrophobic at δ97 (ref. 19). Of the δ-chain residues that form the binding pocket, only Tyr 34 of CDR1δ and Arg 51 of CDR2δ are part of the *Vδ2* gene segment (Fig. 3d). Both of these residues are also found in *Vδ1*. The other residues in the putative binding site derive from *Jδ1* or *Dδ3* gene segments, or are N-encoded. An unusual feature of the Vδ2 domain is an unpaired and almost entirely buried cysteine residue that is located at the base of the CDR3 loop and is not present in other human or murine *Vδ* gene sequences. The putative antigen-binding site described here may explain the broad antigen reactivity by polyclonal populations of Vγ9Vδ2$^+$ T cells, as most residues that seem necessary for phosphoantigen recognition are encoded by either *V* or *J* gene segments.

Like αβTCRs, a functional γδ receptor at the T-cell surface is a complex of the TCR chains and the CD3γ, -δ, -ϵ and -ζ subunits, which associate with cytoplasmic proteins for T-cell signalling. The extracellular portions of the CD3γ, -δ and -ϵ subunits contain a single immunoglobulin-like domain and may interact with the C domains of the TCR. To discern how CD3 interacts with both γδ- and αβTCRs, we looked for a conserved surface to which CD3 may bind. Examination of both the shape and nature of the surfaces of CαCβ and CγCδ revealed no clear similarities. Furthermore, there are only a few solvent-exposed residues that are structurally conserved in both Cβ and Cγ domains. Residues Lys γ186 and Tyr γ191 flank the DE loop on both sides and are near, but not involved with, the Vγ–Cγ interface. Residue Arg γ209, on the opposite side of the Cγ domain in the middle of strand F, is surrounded by other charged residues (Glu γ207, Lys γ171, Lys γ172) that are not found in Cβ. We also compared the location of N-linked glycosylation sites between the C domains. Although the position of the single glycosylation site in Cγ is structurally equivalent to the single site in human Cβ, only one (Asn δ134) of the two sites in Cδ can be regarded as spatially equivalent to one (Asn α183) of the three glycosylation sites in Cα. From this comparison of TCRs, it is not possible to conclude where or how the extracellular domains of CD3 subunits may interact with the extracellular portions of TCRs.

The C domains of the γδTCR are structurally different from those of αβTCRs. First, although sequence alignments predicted that the Cγ domain would have a long, Cβ-like FG loop[21], the structure of G115 reveals Cγ to have a much shorter FG loop (Fig. 4, in red at right; see also Fig. 2 of Supplementary Information. The Cγ FG loop is seven residues in length and is tucked in close to the core of the Cγ domain, whereas the Cβ FG loop is 19 residues in length and extends away from the core of Cβ. As the human and murine Cγ sequences have 76% sequence identity in their FG loops and F and G strands, the murine Cγ domain will probably also possess a short FG loop. The FG loop of Cβ has been proposed to be part of the binding site for the CD3ϵ subunit[22]. The much smaller Cγ loop suggests that, if CD3ϵ binds to Cγ, the FG loop of Cγ is not the primary site of CD3ϵ interaction. The second difference between the C domains of γδ and αβTCRs is that the secondary structure of Cδ is unlike that of Cα. Cδ is composed of a regular immunoglobulin-like domain with a regular three-stranded *β*-sheet as its outer face; Cα contains unusual secondary structure instead of a regular outer *β*-sheet (Fig. 4, in green at left). Clearly, when compared to the αβTCR C domains, the C domains of γδTCRs will present different molecular surfaces to CD3 subunits and to other cell-surface molecules. This could allow one kind of signalling complex to form with γδTCRs and a different kind to form with αβTCRs.

The two kinds of TCRs differ in the number of residues between the end of the C domains and in the position of the interchain disulphide bond. In αβTCR structures, the last ordered residues of Cα and Cβ are 6 or 7 residues and 1 or 2 residues, respectively, away from the interchain disulphide. In G115, the electron density indicates that δ206 and γ230 are the last ordered residues, although the δ- and γ-chains that are in the crystals are longer and end at residues δ229 and γ242, just before the interchain disulphide cysteines. Thus, there are 23 and 12 residues of the δ-chain and γ-chain, respectively, between the end of the C domains and the interchain disulphide. The number of residues between the interchain disulphide and the transmembrane domain is almost the same in both γδTCRs and αβTCRs. Other γδTCRs that use the *Cγ2* gene segment lack an interchain disulphide, and alternative splicing of additional *Cγ2* exons can result in an even longer span between the end of the Cγ domain and the membrane. The longer protein chain between the γδ C domains and the interchain disulphide may provide an additional means of making γδTCRs distinct by means of the altered association of CD3 subunits or by a greater flexibility on the membrane of a γδ receptor.

Structural differences between γδTCRs and αβTCRs both in overall shape and in the C domains suggest a specific role for γδTCRs in immunity and point to those same regions in αβTCRs as being important for αβTCR function. The germline-encoded phosphoantigen-binding site described here may account for the broad antigen reactivity of polyclonal, peripheral blood γδ$^+$ T cells that can be activated simultaneously by similar ligands, producing a significant effect on the immune system. □

Methods

Crystallization

The G115 (ref. 3) (or G9 (refs 8, 19, 23)) γ- and δ-chains (Vγ9JPC1 and Vδ2D3J1C, see http://imgt.cines.fr) were expressed as insoluble proteins in *Escherichia coli* with and without selenomethionine, then refolded together by rapid dilution in 1 l of 1 M L-arginine, 0.1 M Tris-HCl, pH 8.0 and 0.2 mM reduced/0.2 mM oxidized glutathione. After dialysis against 10 mM Tris-HCl pH 8.0, concentration by cation exchange chromatography at pH 5.5, and purification by size exclusion chromatography at pH 8.0, the refolded protein formed crystals in hanging drops containing 4 mg ml^{-1} protein, 28% polyethylene glycol 4000, 0.2 M Li_2SO_4, 0.1 M Tris-HCl, pH 8.5 and 20% glycerol.

Structure determination

X-ray data were measured at three wavelengths on a selenomethionine-containing crystal at beamline X9B of the National Synchrotron Light Source (Brookhaven National Laboratory) and at the CuKα wavelength (525° of data) on a crystal soaked in the synthetic phosphoantigen bromohydrin pyrophosphate (BrHPP)[18]. Data were integrated and scaled with HKL2000 (ref. 24). We located selenium atoms (12 out of 28) using SnB[25]. Refinement of selenium positions and calculation of experimental phases were performed with SHARP[26]; values of f″ were each refined as single-wavelength anomalous diffraction calculations, followed by f′ refinement as a multiwavelength anomalous diffraction (MAD) calculation. Four sets of selenium atoms (20 total) exhibited non-crystallographic symmetry (NCS). Density modification, performed with DM from the CCP4 suite[27], confirmed that there were four γδTCR molecules present in the asymmetric unit. Iterative cycles of NCS averaging, reciprocal space phase combination, and other crystallographic calculations were performed with the RAVE[28] and CCP4 (ref. 27) program packages. Crystallographic statistics are presented in Supplementary Information.

The G115 model was built using O (ref. 29) and refined against the MAD data using the maximum likelihood Hendrickson–Lattman (MLHL) target in CNS[30]. The nearly complete model was refined against the higher resolution data from the antigen-soaked native crystal using the maximum likelihood amplitude (MLF) target in CNS. Small crystals (< 100 μm) and high redundancy contributed to poor merging statistics; however, intensities for the highest bin of resolution were well over 2σ and R_{free} for the highest resolution bin was less than 40% for the final model. Therefore, all data to 3.12 Å were included in refinement. The anomalous signal of these data (with model phases) was able to confirm the locations of sulphur atoms at cysteines, methionines and two sulphates involved in a crystal contact, but did not reveal any sites that could be attributed to the brominated antigen. Furthermore, the presence of the antigen could not be confirmed from the electron density map. Tight NCS restraints were applied during initial model refinement, but were later relaxed to 200 kcal mol^{-1} Å^{-2} for main-chain atoms and 100 kcal mol^{-1} Å^{-2} for side-chain atoms. Each of the four domains of the TCR was restrained by a separate NCS operator, thus allowing the movement of domains relative to each other. The r.m.s. deviations between equivalent domains are less than 0.05 Å and 0.09 Å for main-chain and side-chain atoms, respectively. NCS restraints were not applied to the CDRs, switch regions and loops that exhibited modest conformational variation between molecules.

R_{free} was monitored throughout refinement using reflections (5% of total) that were selected in 35 resolution shells. Water molecules were picked using the water_pick and water_delete scripts in CNS into locations that exhibited 3.5σ density in an F_o–F_c map; 1.3σ density in a 2F_o–F_c map; 1σ density in a simulated anneal 2F_o–F_c composite omit map; and proper hydrogen-bonding distance and geometry and that were visually confirmed in the program O. The final model (R = 21.9% and R_{free} = 26.6%) includes four copies of residues δ1–206 and γ1–230, plus 54 water molecules and 2 sulphates in the asymmetric unit. A Ramachandran plot of the model shows 81% of residues in the most favoured regions and 17% in additionally allowed regions, with no residues in disallowed

© 2001 Macmillan Magazines Ltd

regions. Amino-acid side chains exhibiting weak electron density were built with steric and geometric constraints. We generated the figures with GRASP, Molscript and Raster3D.

Structural comparisons

We carried out structural comparisons against 251 Fab and 11 αβTCR models. Superpositions were performed using LSQMAN[28]. The inter-domain, intra-chain angles (for instance Vγ–Cγ) were calculated as the supplement of the rotation angle necessary for superposition of the conserved α-carbon atoms of the B and F strands of each domain. Because the Vγ–Vδ interface involves the A′GFCC′C″ β-sheet whereas the Cγ–Cδ interface involves the ABED β-sheet (Fig. 1), V domain strands B and F were explicitly superimposed onto C domain strands F and B. The elbow angle between the V and C domains was calculated as the angle between the pseudo-dyads, described by direction cosines, which result from a superposition of V or C domains. Buried interfaces were calculated with CNS using a probe radius of 1.4 Å. Structures of αβTCRs and Fabs are referred to by their Protein Data Bank identification codes (http://www.rcsb.org).

Received 16 January; accepted 30 April 2001.

1. Hayday, A. C. γδ cells: a right time and a right place for a conserved third way of protection. *Annu. Rev. Immunol.* **18,** 975–1026 (2000).
2. Chien, Y. H., Jores, R. & Crowley, M. P. Recognition by γ/δ T cells. *Annu. Rev. Immunol.* **14,** 511–532 (1996).
3. Constant, P. *et al.* Stimulation of human γδ T cells by nonpeptidic mycobacterial ligands. *Science* **264,** 267–270 (1994).
4. Rock, E. P., Sibbald, P. R., Davis, M. M. & Chien, Y. H. CDR3 length in antigen-specific immune receptors. *J. Exp. Med.* **179,** 323–328 (1994).
5. Sciammas, R. *et al.* Unique antigen recognition by a herpesvirus-specific TCR-γδ cell. *J. Immunol.* **152,** 5392–5397 (1994).
6. Morita, C. T. *et al.* Direct presentation of nonpeptide prenyl pyrophosphate antigens to human γδ T cells. *Immunity* **3,** 495–507 (1995).
7. Lang, F. *et al.* Early activation of human Vγ9Vδ2 T cell broad cytotoxicity and TNF production by nonpeptidic mycobacterial ligands. *J. Immunol.* **154,** 5986–5994 (1995).
8. Davodeau, F. *et al.* Close correlation between Daudi and mycobacterial antigen recognition by human γδ T cells and expression of V9JPC1γ/V2DJCδ-encoded T cell receptors. *J. Immunol.* **151,** 1214–1223 (1993).
9. Fournié, J. J. & Bonneville, M. Stimulation of γδ T cells by phosphoantigens. *Res. Immunol.* **147,** 338–347 (1996).
10. Belmant, C. *et al.* 3-formyl-1-butyl pyrophosphate: a novel mycobacterial metabolite-activating human γδ T cells. *J. Biol. Chem.* **274,** 32079–32084 (1999).
11. Tanaka, Y., Morita, C. T., Nieves, E., Brenner, M. B. & Bloom, B. R. Natural and synthetic non-peptide antigens recognized by human γδ T cells. *Nature* **375,** 155–158 (1995).
12. Bukowski, J. F., Morita, C. T. & Brenner, M. B. Human γδ T cells recognize alkylamines derived from microbes, edible plants, and tea: implications for innate immunity. *Immunity* **11,** 57–65 (1999).
13. Kunzmann, V. *et al.* Stimulation of γδ T cells by aminobisphosphonates and induction of antiplasma cell activity in multiple myeloma. *Blood* **96,** 384–392 (2000).
14. Li, H. *et al.* Structure of the Vδ domain of a human γδ T-cell antigen receptor. *Nature* **391,** 502–506 (1998).
15. Loh, E. Y. *et al.* Gene transfer studies of T cell receptor-γδ recognition. Specificity for staphylococcal enterotoxin A is conveyed by Vγ9 alone. *J. Immunol.* **152,** 3324–3332 (1994).
16. Li, H. *et al.* Three-dimensional structure of the complex between a T cell receptor β chain and the superantigen staphylococcal enterotoxin B. *Immunity* **9,** 807–816 (1998).
17. Morita, C. T. *et al.* Superantigen recognition by γδ T Cells: SEA recognition site for human Vγ2 T cell receptors. *Immunity* **14,** 331–344 (2001).
18. Belmant, C. *et al.* A chemical basis for selective recognition of nonpeptide antigens by human γδ T cells. *FASEB J.* **14,** 1669–1670 (2000).
19. Davodeau, F. *et al.* Peripheral selection of antigen receptor junctional features in a major human γδ subset. *Eur. J. Immunol.* **23,** 804–808 (1993).
20. Panchamoorthy, G. *et al.* A predominance of the T cell receptor Vγ2/Vδ2 subset in human mycobacteria-responsive T cells suggests germline gene encoded recognition. *J. Immunol.* **147,** 3360–3369 (1991).
21. Bentley, G. A., Boulot, G., Karjalainen, K. & Mariuzza, R. A. Crystal structure of the β chain of a T cell antigen receptor. *Science* **267,** 1984–1987 (1995).
22. Wang, J. *et al.* Atomic structure of an αβ T cell receptor (TCR) heterodimer in complex with an anti-TCR Fab fragment derived from a mitogenic antibody. *EMBO J.* **17,** 10–26 (1998).
23. Davodeau, F. *et al.* Secretion of disulfide-linked human T-cell receptor γδ heterodimers. *J. Biol. Chem.* **268,** 15455–15460 (1993).
24. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276,** 307–326 (1997).
25. Smith, G. D., Nagar, B., Rini, J. M., Hauptman, H. A. & Blessing, R. H. The use of SnB to determine an anomalous scattering substructure. *Acta Crystallogr. D* **54,** 799–804 (1998).
26. de la Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods. *Methods Enzymol.* **276,** 472–494 (1997).
27. Bailey, S. The CCP4 suite—programs for protein crystallography. *Acta Crystallogr D* **50,** 760–763 (1994).
28. Kleywegt, G. J. & Jones, T. A. in *From First Map to Final Model* (eds. Bailey, S., Hubbard, R. & Waller, D.) 59–66 (SERC Daresbury Laboratory, Warrington, 1994).
29. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for binding protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
30. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).

Supplementary information is available on *Nature's* World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank C. Hammer for mass spectrometry; M. Garfield for amino-acid sequencing; S. Garman for advice and discussions; and Z. Dauter and K. R. Rajashankar for help at beamline X9B at the National Synchrotron Light Source at Brookhaven National Laboratory, which is supported by the US Department of Energy, Division of Materials Sciences and Division of Chemical Sciences. T.J.A. is supported by a postdoctoral fellowship from the Cancer Research Institute. This work is supported by the intramural program of the National Institute of Allergy and Infectious Diseases.

Correspondence and requests for materials should be addressed to T.J.A. (e-mail: tallison@niaid.nih.gov) or D.N.G. (e-mail: garboczi@nih.gov).
Coordinates have been deposited at the Protein Data Bank (identification code 1hxm) and will be released on publication.

corrections

Plant diversity enhances ecosystem responses to elevated CO_2 and nitrogen deposition

Peter B. Reich, Jean Knops, David Tilman, Joseph Craine, David Ellsworth, Mark Tjoelker, Tali Lee, David Wedin, Shahid Naeem, Dan Bahauddin, George Hendrey, Shibu Jose, Keith Wrage, Jenny Goth & Wendy Bengston

Nature **410,** 809–812 (2001).

This paper should have included the following acknowledgement:

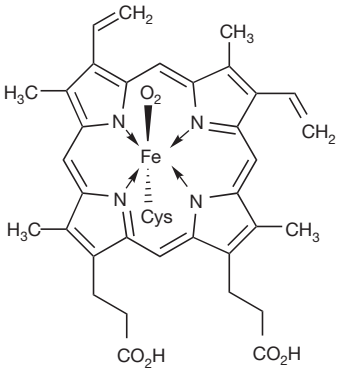
We thank the US Department of Energy for financial support, with additional support from the US National Science Foundation Long-Term Ecological Research Program. □

Enabling the chemistry of life

C. Walsh

Nature **409,** 226–231 (2001).

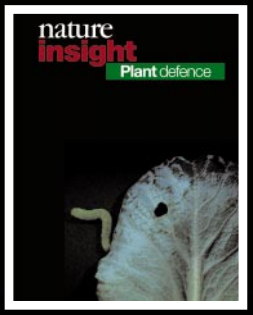
In Fig. 4 of this *Insight* article, the P450 haem active site is wrongly shown as a uroporphyrinogen I rather than a protoporphyrin IX macrocycle. The correct version of the molecule is shown here. □



© 2001 Macmillan Magazines Ltd

nature insight

Plant defence



Cover illustration courtesy of P. Reymond and E. Farmer.

This is an exciting time to be interested in plant science. The advances in plant molecular biology, of which the sequencing of the *Arabidopsis thaliana* genome is but the most visible manifestation, are providing revelations on a weekly basis, and one of the faster moving fields within plant science is the study of plant defence mechanisms.

Many of the genetic and physiological foundations of the study of plant defence were laid back in the early part of the twentieth century, but it was only seven years ago that the first gene concerned with plant disease resistance was cloned. From a purely academic standpoint this is a 'road less travelled'. While many immunologists may view the mammalian immune system as a pinnacle of evolution, natural selection has honed the defence strategies of plants over 1.6 billion years without recourse to antibodies, T cells and the like, producing systems no less subtle or effective.

One theme that emerges in this Insight is how molecules and mechanisms involved in plant defence have direct homologues in animals. At least some of the plant resistance genes (*R* genes) considered by Dangl and Jones (pages 826–833) are cousins of the Toll-like receptors whose importance in animals' innate immunity has recently been recognized. The phenomenon of post-transcriptional gene silencing, which Waterhouse, Wang and Lough (pages 834–842) present as a defence against viruses and transposons, seems practically identical to RNAi discovered nearly simultaneously in *Caenorhabditis elegans*. Programmed cell death, used as a way to isolate and excise diseased plant tissue, is shown by Lam, Kato and Lawton (pages 848–853) to proceed in ways both similar and different to the induction of apoptosis in animals. Even pheromones have their counterparts in the volatile compounds used in inter- and intra-plant communication, and for which Farmer (pages 854–856) coins the term 'automone'. But despite the similarities, these systems have evolved in directions and to levels of sophistication not seen in animals.

A knowledge of plant defence strategies also has practical applications. The 'natural' products of plants' secondary metabolism have been used for millennia in 'traditional medicines', but the function of most of these elaborate chemicals is to protect the plant from attack. Dixon (pages 843–847) surveys the diversity of plants' chemical warfare on pathogens and considers the potential for metabolic engineering of natural product pathways, while Stuiver and Custers (pages 865–868) explore how our current knowledge of plant defence mechanisms is being, or soon will be, exploited to produce improved disease resistance in crops. But, as Rausher points out (pages 857–864), such things are not new. Evolution has been introducing new defence strategies into plants for millions of years while simultaneously finding ways for pathogens to bypass them. We will need to learn lessons from ecology as well as biochemistry if we are to apply our emerging understanding of plant defence in any meaningful way.

Seven articles is too few to provide anything but a survey of some aspects of this fascinating subject. However, we hope that this Insight will give readers a better idea of the armaments employed for the battle raging in our own backyards.

Christopher Surridge Senior Editor

Lesley Anson Insight Programme Editor

© 2001 Macmillan Magazines Ltd

Plant pathogens and integrated defence responses to infection

Jeffery L. Dangl* **& Jonathan D. G. Jones†**

**Department of Biology and Curriculum in Genetics, Coker Hall 108, CB#3280, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3280, USA (e-mail: dangl@email.unc.edu)*
†The Sainsbury Laboratory, John Innes Centre, Colney, Norwich NR4 7UH, UK (e-mail: jonathan.jones@bbsrc.ac.uk)

Plants cannot move to escape environmental challenges. Biotic stresses result from a battery of potential pathogens: fungi, bacteria, nematodes and insects intercept the photosynthate produced by plants, and viruses use replication machinery at the host's expense. Plants, in turn, have evolved sophisticated mechanisms to perceive such attacks, and to translate that perception into an adaptive response. Here, we review the current knowledge of recognition-dependent disease resistance in plants. We include a few crucial concepts to compare and contrast plant innate immunity with that more commonly associated with animals. There are appreciable differences, but also surprising parallels.

Most plants are resistant to most plant pathogens. Passive protection against pathogens that are not specialized to attack a specific host is provided by waxy cuticular 'skin' layers and preformed anti-microbial compounds. Plant pathogens can be broadly divided into those that kill the host and feed on the contents (necrotrophs) and those that require a living host to complete their life cycle (biotrophs). Microbial necrotrophy is often accompanied by production of toxins. Viruses are quintessential biotrophs, although infection can lead eventually to host cell death. Bacteria and fungi can adopt either lifestyle. Many insects cause damage by chewing. They induce a wound response that includes the production of protease inhibitors and other anti-feedants such as alkaloids. Additionally, wound responses include release of volatiles which attract insects that feed on, or deposit eggs into, the larvae of the herbivorous insect. By contrast, sap-feeding insects and nematodes can adopt more intimate and sophisticated modes of biotrophic parasitism, imposing developmental responses on the plant cells, leading to the appearance of galls, root knots or cysts. The plant innate immune response is highly polymorphic in its capacity to recognize and respond to biotrophs, and we focus here on this aspect of plant defence.

Resistance in hosts and avirulence in pathogens

Plant–pathogen interactions, particularly those involving biotrophic parasites, are governed by specific interactions between pathogen *avr* (avirulence) gene loci and alleles of the corresponding plant disease resistance (*R*) locus. When corresponding *R* and *avr* genes are present in both host and pathogen, the result is disease resistance. If either is inactive or absent, disease results[1]. The simplest model that accounts for this genetic interaction requires that *R* products recognize *avr*-dependent signals and trigger the chain of signal-transduction events that culminates in activation of defence mechanisms and an arrest of pathogen growth. *R* genes specify a polymorphic component of a particular recognition event. Specific *R*-mediated innate immunity is superimposed onto one or more basal defence pathways. Basal defences inhibit pathogen spread after successful infection and onset of disease. The existence of basal defence is inferred from the identification of mutants that are more susceptible to a virulent pathogen than are their parents (detailed below). Genetic overlap between specific and basal resistance responses suggests that one function of *R*-mediated signalling is to more rapidly and effectively activate defence mechanisms that are shared by both pathways[2–4].

A significant effort by several laboratories in the past 5–10 years has resulted in the identification of many *R* genes from model and crop species[5–7]. Functional *R* genes isolated so far encode resistance to bacterial, viral, fungal, oomycete and even nematode and insect pathogens with very different lifestyles, outside or inside the plant cell. Despite this wide range of pathogen taxa and their presumed pathogenicity effector molecules, *R* genes encode only five classes of proteins (Fig. 1).

The largest class of *R* genes encodes a 'nucleotide-binding site plus leucine-rich repeat' (NB-LRR) class of proteins (Fig. 1). These function, so far, exclusively as *R* genes and they are highly evolved (see below) for that function. Although computer analyses do not predict localization, at least one NB-LRR protein is associated with the plasma membrane[8]. Their most striking structural feature is a variable number of carboxy-terminal LRRs. LRR domains are found in diverse proteins and function as sites of protein–protein interaction, peptide–ligand binding and protein–carbohydrate interaction[9,10]. In addition, each R protein contains a conserved nucleotide-binding (NB) site, which in other proteins is critical for ATP or GTP binding[11]. But it is not clear how or which of these nucleotides is bound. The nucleotide-binding site is part of a larger domain that includes additional homology between R proteins and some eukaryotic cell death effectors such as Apaf-1 and Ced4 (Fig. 2). This enlarged region is termed the NB-ARC or Ap-ATPase domain[12,13]. By analogy with Apaf-1 function, activation of R proteins may involve Avr-dependent release of the Ap-ATPase domain from inhibition by the C-terminal LRRs, followed by multimerization of a complex that recruits additional proteins to the amino-terminal domain for further signalling events. The NB-LRR class can be subdivided based on deduced N-terminal structural features: many have a domain with homology to the intracellular signalling domains of the *Drosophila* Toll and mammalian interleukin (IL)-1 receptors (TIR-NB-LRR),

© 2001 Macmillan Magazines Ltd

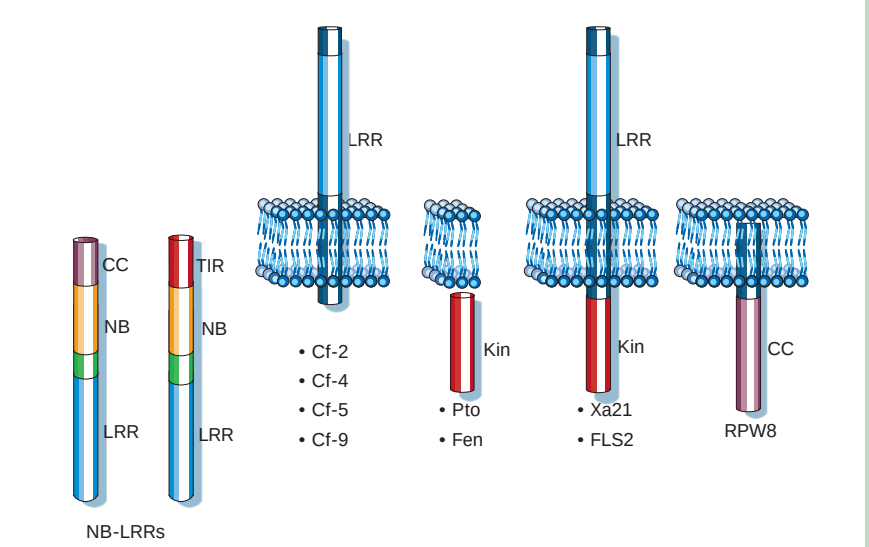


Figure 1 Representation of the location and structure of the five main classes of plant disease resistance proteins. Xa21 and Cf-X proteins carry transmembrane domains and extracellular LRRs. The recently cloned RPW8 gene product carries a putative signal anchor at the N terminus. The *Pto* gene encodes a cytoplasmic Ser/Thr kinase, but may be membrane associated through its N-terminal myristoylation site. The largest class of R proteins, the NB-LRR class, are presumably cytoplasmic (although they could be membrane associated) and carry distinct N-terminal domains.

whereas others contain putative coiled-coil domains (CC-NB-LRR). The CC-NB-LRR class probably comprises multiple subfamilies, varying in size and in the location of the coiled-coil domain.

Comparative sequence analyses demonstrated that *R* specificity resides largely in the LRRs, which are under diversifying selection to increase amino-acid variability in residues thought to be solvent exposed[14–18]. Construction of domain chimaeras has supported these findings for both NB-LRR and extracellular LRR classes of R proteins[19–22]. Recent evidence indicates that in the *L* class of flax rust resistance genes, diversifying selection also acts on residues in the TIR domain, and that these residues are apparently co-evolving with the corresponding LRR domain to provide specificity[23]. Mechanisms for the evolution of new specificities include unequal recombination and gene conversion, as well as accumulation of amino-acid codon exchanges in members of anciently duplicated gene families.

R-gene diversity

The complete *Arabidopsis* sequence permits a comprehensive analysis of the diversity of NB-LRR *R*-gene sequences in one plant[24]. Annotation revealed ~150 sequences with homology to the NB-LRR class of *R* genes. *R* homologues are unevenly distributed between chromosomes, with 49 on chromosome 1, 2 on chromosome 2, 16 on chromosome 3, 28 on chromosome 4, and 55 on chromosome 5. Not all of these seem intact. Despite the fact that many previously isolated *R* genes seem to reside in local multigene families, there are 46 singleton *Arabidopsis R*-gene homologues, 25 doublets, 7 loci with three copies, and individual loci with four, five, seven, eight and nine NB-LRR-encoding genes. In recent months, the *RPP7* family has been defined as an additional cluster of ~14 copies on chromosome 1 (A. Cuzick and E. Holub, personal communication). A continuously updated annotation of *Arabidopsis R* genes by B. Meyers and colleagues can be found at http://pgfsun.ucdavis.edu/niblrrs/At_RGenes/. There are more TIR-NB-LRR genes (~60%) than CC-NB-LRR genes (~40%). Both inverted and direct repeats of *Arabidopsis R* genes exist, at a ratio of about 3:2. The largest clusters are at the *RPP5* locus, which carries the *RPP4* gene and seven other *RPP5* homologues, and at the complex *RPP7* locus on chromosome 1.

Thus, *Arabidopsis* has ~100 *R* loci distributed over all the chromosomes. This seems a surprisingly small number of genes to mediate recognition of all possible pathogen-encoded ligands. Several models could explain this. Perhaps many R proteins actually perceive the presence of more than one Avr protein, whether that Avr protein comes from pathogens of similar or different lifestyles. 'Dual recognition' has been demonstrated in a few cases. For example, *RPM1* recognizes two non-homologous *avr* genes[25,26], the tomato *Mi* gene confers not only nematode resistance but also aphid resistance[27], and alleles of the *RPP8/HRT* gene recognize an oomycete parasite and a virus[15,28]. Similarly, the closely related potato *Rx* and *Gpa2* genes confer virus and nematode resistance, respectively[29]. Alternatively, it is possible that some R proteins recognize conserved pathogen molecules, and are of ancient origin[30,31]. If this is the case, then it is plausible that stable polymorphism for ancient *R*-gene specificity is important for restricting disease in wild populations. Furthermore, one locus can evolve to generate an allelic series that can confer recognition capacity of multiple *avr* genes. Polymorphism for recognition capacity will be sustained by frequency-dependent selection, provided that polymorphism for avirulence is present in the pathogen population[32].

Of considerable interest is the identification of truncated forms of both CC-NB and TIR-NB genes that lack the LRRs. It remains to be determined whether these are simply the unpurged debris of past mutation events, or whether they encode adaptor molecules that are important in signalling, as *MyD88* contributes to TIR signalling in animals[33]. *MyD88* encodes a protein with a TIR domain and a death domain that recruits IL-1 receptor-associated kinase (IRAK) to Toll-like receptors or the IL-1 receptor. Alternatively spliced versions of the TIR-NB-LRR proteins N and L are observed, and may have as yet unidentified roles in disease resistance[34]. There are also some unexpected structures. Two genes encode, in addition to a TIR-NB-LRR structure, a WRKY domain that is likely to confer DNA-binding capacity. WRKY proteins are plant-specific zinc-finger transcription factors that are transcriptionally activated during some plant defence responses[35]. In addition, one TIR-NB-LRR gene has been annotated to carry not only a WRKY, but also a protein kinase domain.

The Col-0 genome sequence represents a single, inbred haplotype, and comparison to other inbreds (termed accessions) will require additional work. For example, *RPM1* is absent from accession Nd-0 (refs 26, 36). Conversely, there may be Nd-0 *R* genes that are lacking from Col-0, such as the newly isolated *RPW8* class of *R* gene (see later). The *RPP8* gene is single copy in Col-0, and duplicated in La-*er*[37], and the *RPP4/5* haplotypes in Col-0 and La-er are extremely diverged[38]. Extensive analysis by DNA sequencing and gel blot hybridization of homologues from multiple accessions will provide further insights into *R*-gene evolution.

The other four classes of *R* genes are structurally diverse (Fig. 1). In addition, some members of these gene families have demonstrated functions in cellular and developmental processes unrelated to defence. *Pto* from tomato encodes a Ser/Thr kinase that confers resistance to *Pseudomonas syringae* strains carrying *avrPto*. Pto might function through a phosphorylation cascade, triggered by direct

© 2001 Macmillan Magazines Ltd

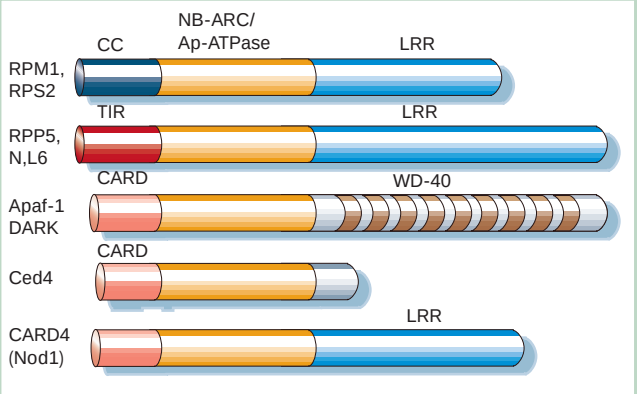


Figure 2 Comparison of R proteins with proteins involved in cell death in animal cells. In addition to the nucleotide-binding site, Apaf-1, R proteins and Ced4 carry further homologies in the NB-ARC domain that might suggest similarity in mode of action[12,13]. In Apaf-1, the WD-40 repeats seem to confer negative regulation of cell death that is relieved on binding cytochrome *c*, resulting in CARD domain multimerization, caspase-3 recruitment and apoptosis. Conceivably, the R proteins undergo Avr product-dependent multimerization of their N-terminal domains and recruitment of additional signalling proteins.

AvrPto–Pto interaction[39,40]. Pto function requires the NB-LRR protein Prf[41]. The rice *Xa21* gene encodes a transmembrane receptor carrying a large extracellular LRR domain and an intracellular protein kinase domain[42]. Chimaeras of Xa21 and a related LRR receptor-like kinase that recognizes the brassinosteroid hormone show that specificity for this class of R protein also resides in the LRRs[43,44]. The tomato *Cf-X* genes encode single pass membrane proteins with extracellular LRRs[45]. These last two structural classes are reminiscent of the *Arabidopsis CLV1* and *CLV2* genes, which may function together to recognize an extracellular peptide ligand encoded by *CLV3* (ref. 46). Intriguingly, *FLS2*, a gene required for *Arabidopsis* to recognize a conserved amino-acid sequence present in bacterial flagellin, also encodes a *CLV1/Xa21* homologue[47]. A new *R* gene in *Arabidopsis* (*RPW8*; ref. 37) encodes a small, probable membrane protein with a possible coiled-coil domain and essentially no other homology to known proteins.

Whether these other structural classes of R proteins use signal-transduction cascades similar to those used by the NB-LRR family is not yet known, although three findings suggest they do. First, *Prf* functions with *Pto*, as noted above. Second, the *Arabidopsis PBS1* gene (required for the function of the *RPS5* NB-LRR gene, but not the related *RPM1* and *RPS2* genes) is also a Ser/Thr kinase, suggesting that these two classes of proteins may often function together in *R* signal-transduction pathways[48]. There are over 50 *Arabidopsis* protein kinase genes that are strongly homologous to *Pto*. Third, *RPW8* activity, and the activity of several TIR-NB-LRR proteins, is dependent on *EDS1* (ref. 37). *Arabidopsis* also carries homologues of other *R*-gene classes, including 174 homologues of the *Xa21* class of LRR receptor-like kinase. There are also 30 genes that resemble *Cf-9*, or *CLV2*, in that they encode extracellular LRRs and a short cytoplasmic domain. The nature of signal-transduction cascades downstream of activation of these classes is so far unknown. Whether any of these function as the polymorphic component in pathogen recognition, or in concert with NB-LRR proteins, remains to be determined.

LRR receptor polymorphism in animal innate immunity

Animal innate immune systems also use LRR receptors, of the extracellular variety, called Toll-like receptors or TLRs (named after the first member identified; Fig. 3). The mammalian and *Drosophila* innate immune receptors couple to internal cell-death signals, kinase cascades and effector arms that are transcriptionally activated (reviewed in refs 49, 50). There are probably ~15–20 TLR genes in the human genome, and perhaps twice that number of TIR domain-containing proteins (D. Golenbock, personal communication). TLRs recognize a limited, but highly conserved and common, set of pathogen-encoded structures that may represent signatures for a given pathogen class. Recent reports of combinatorial functions of TLR proteins suggest a modest expansion of the germline repertoire[51]. However, it seems that the overall recognition potential is limited, albeit to important and non-mutable ligands such as lipopolysaccharide (LPS).

In addition, there is a class of intracellular NB-LRR proteins that also have a role in animal innate immunity. Surprisingly, these are structurally analogous to the NB-LRR class of plant R proteins. These mammalian Nod proteins contain N-terminal caspase-activating recruitment domains (CARDs) and NB-ARC/Ap-ATPase domains, like Apaf-1. But the Nod proteins also carry C-terminal LRRs, like plant R proteins, instead of WD-40 repeats, like Apaf-1. Intriguingly, Nod1 (also called CARD4) confers recognition of bacterial LPS and subsequent NF-κB activation in a TLR4-independent manner[52]. Although Nod1 is broadly expressed in many cell types, Nod2 is expressed primarily in monocytes, a key cell type that binds and can engulf bacteria in the animal innate immune system[53]. NF-κB activation following Nod2 stimulation occurs probably through a direct interaction with the RICK Ser/Thr protein kinase, mediated by the Nod2 CARD domains[53]. Most important, mutations in *Nod2* have recently been implicated in Crohn's disease[54,55], an inflammatory bowel disorder that phenotypically resembles an autoimmune disease. The LRR domains of both Nod1 and Nod2 are required for function. The existence of ~30 *Nod* genes with similar NB-LRR structures, but possessing varied N-terminal domains, suggests mammals may carry a system of intracellular receptors that, like plant R proteins, determines recognition of intracellular ligands. There is currently no evidence that polymorphism in either the TLR or Nod proteins further contributes to diversity. It could be that subsequent to the evolution of the adaptive immune system, there was no evolutionary pressure for expansion of TLRs in animals as they were already adapted to recognize key, non-mutable pathogen-encoded ligands.

Complex evolution

If several proteins in the recognition and response pathway are functionally polymorphic, then the optimal set of proteins will need to evolve together. The existence of cytoplasmic and transmembrane classes of R protein indicates that some are specialized to detect secreted ligands or surface components from the pathogen, and some are dedicated to recognize ligands that appear inside the cell (see below). The discovery of multiple types of both intracellular and transmembrane R proteins suggests that polymorphism may exist not only in recognition but also in several elements in response pathways. In plant breeding, this polymorphism may be uncovered experimentally in simple pairwise comparisons of resistant and susceptible inbred host lines. Yet it is probable that selective pressure is not acting on only a single pathogen recognition element, assuming that the LRRs are either directly or indirectly responsible for ligand contact. Selection could act to diversify and then fine-tune the output of the response. NB-LRR proteins probably work in complexes. Thus, it may be that genetic buffering[56], whereby evolutionary experiments in polymorphism are protected by redundancy, and by flexible, yet robust signalling processes, facilitates phenotypic variation. This, in turn, allows a flexible evolutionary space in which to diversify several elements of the system.

Because most proteins work in complexes, it is perhaps unsurprising that co-evolution of the components is required for optimization. This strategy is most easily detected in outbred populations, and can be revealed as quantitative trait differences among inbred species. Hamilton *et al.*[57] proposed that the main selective advantage leading to the retention of sexual reproduction and

© 2001 Macmillan Magazines Ltd

outcrossing is that polymorphism at loci contributing to parasite recognition will restrict loss of fitness due to disease. According to this model, if a host population is extremely heterogeneous in its recognition capacity, then most isolates of the parasite will not be able to grow on most hosts. In the absence of outcrossing, such polymorphism would be more likely to be lost, unless it is maintained by selection (Fig. 4). Furthermore, if sexual recombination between parasites leads to exchange of dominant avirulence genes, then most progeny of most parasites will not be able to find a host. There is still debate about whether such frequency-dependent selection, in which rare resistance (recognition) specificities are less likely to be overcome by the parasite, is the main explanation for the enormous diversity of human haplotypes at major histocompatibility complex (MHC) loci. An alternative model proposes that this diversity can be explained by overdominance (heterozygote advantage), through which heterozygotes have twice the recognition capacity and resistance of any homozygote. Many plant species, including *Arabidopsis*, reproduce by self-fertilization, and overdominance cannot explain the extreme polymorphism of *R* loci compared with other loci in such species. This inference is of general significance, because it implies that frequency-dependent selection could be part of the explanation for MHC polymorphism in animals, and is perhaps even sufficient to explain it. The restriction of parasite success in plant varietal mixtures is consistent with this overall concept, and the approach deserves further exploration as a strategy to provide more durable resistance in crop varieties (Fig. 4).

NB-LRR proteins are probably intracellular and are likely to be receptors for an *avr*-encoded ligand, or to function in a protein complex that is the functional receptor. Although a cytoplasmic location of an R protein is unsurprising for those conferring resistance to viral pathogens, the existence of intracellular NB-LRR R proteins active against microbial pathogens implies that the ligands from bacterial and fungal pathogens are also intracellular. Plant and animal bacterial pathogens, like *P. syringae*, use a type III delivery system to traffic proteins into host cells (reviewed in refs 58, 59). Avr-R recognition for several bacterial systems (reviewed in ref. 60) occurs inside the plant cell following expression of *avr* genes using plant transcriptional control signals. Curiously, expression of bacterial Avr proteins in disease-susceptible plants can lead to delayed, weak cytotoxic effects, suggesting that Avr proteins may have additional targets inside disease-susceptible plant cells[61–63]. Based on analogy to mammalian pathosystems, it is inferred that type III effectors from phytopathogenic bacteria are translocated into the host cell, although direct demonstrations of this are rare[64]. Despite these recent advances, little is known about how the subcellular localization, and site of action, of type III effector proteins influences initiation of disease, or resistance, in hosts of the appropriate genotype[65].

Many fungal pathogens form intimate membrane contacts with host cells at the surface of specialized feeding structures called haustoria, which could facilitate the traffic of disease effectors into the host. It will be of great significance to isolate *avr* genes from haustorial pathogens, such as rusts, powdery mildews and downy mildews. Several secreted Avr proteins from *Cladosporium fulvum* have been defined that trigger resistance mediated by transmembrane R proteins of the Cf-X family[66–69]. However, they do not seem to interact directly with their corresponding Cf-X proteins. Indeed, Avr9 peptide binds with 70-picomolar affinity to a plasma membrane protein that is present even in tomato lines that lack Cf-9 (ref. 70). The rice blast (*Magnaporthe grisea*) *avrPita* gene has been isolated; it encodes a putative secreted metalloprotease. The corresponding R protein (Pita) is of the CC-NB-LRR class (although the LRRs are rather degenerate), and AvrPita and Pita interact directly in yeast two-hybrid and *in vitro* experiments[71]. Although *M. grisea* is not a haustorial pathogen, this suggests that fungal and downy mildew Avr proteins might be secreted proteins that are internalized by or translocated into the plant cell and recognized intracellularly.

R proteins as guards of cellular machinery

The 'guard hypothesis' provides an intriguing conceptual framework for the action of disease effectors and the R-protein complex. It was put forward in an attempt to rationalize why Pto protein kinase requires the NB-LRR protein Prf to activate defence upon recognition of AvrPto[72]. According to this model, Pto is a general component of host defence, perhaps in a pathway for response to nonspecific elicitors of phytopathogenic bacteria. The function of AvrPto for *P. syringae* is to target Pto and suppress this nonspecific

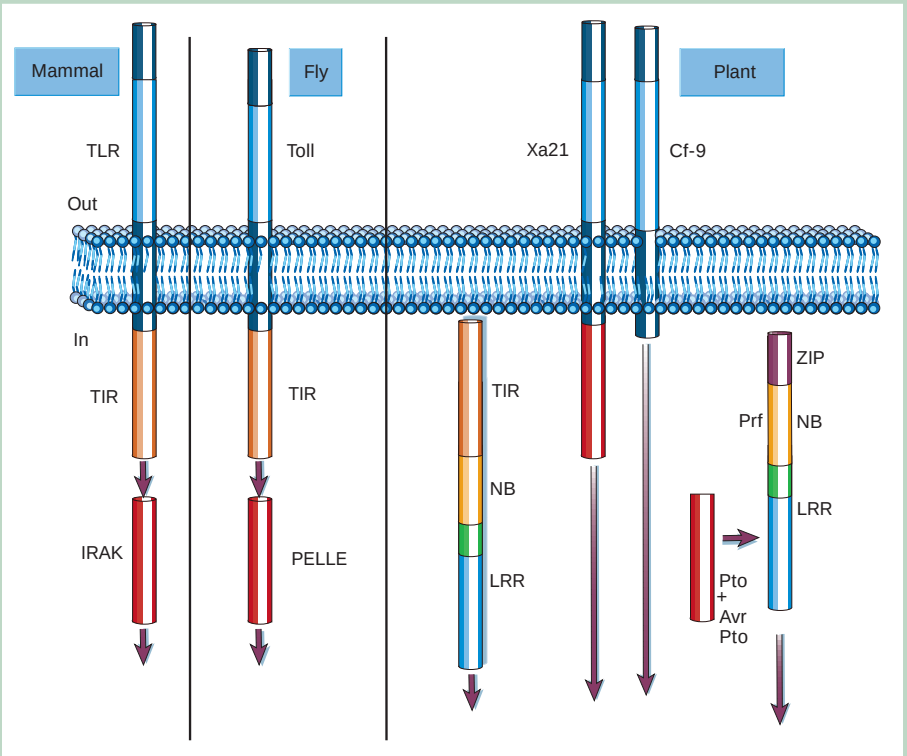


Figure 3 Comparison of R proteins with proteins involved in animal innate immunity. Toll in *Drosophila*, and TLRs in mammals, are activated by extracellular ligands. They carry extracellular LRRs (such as Xa21 and Cf-9), but also carry a cytoplasmic TIR domain. Upon activation, the TIR domain recruits death domain-containing adaptor proteins, and the protein kinases PELLE or IRAK. PELLE and IRAK are the animal proteins that are most homologous to Pto. Pto requires the NB-LRR protein Prf to function.

© 2001 Macmillan Magazines Ltd

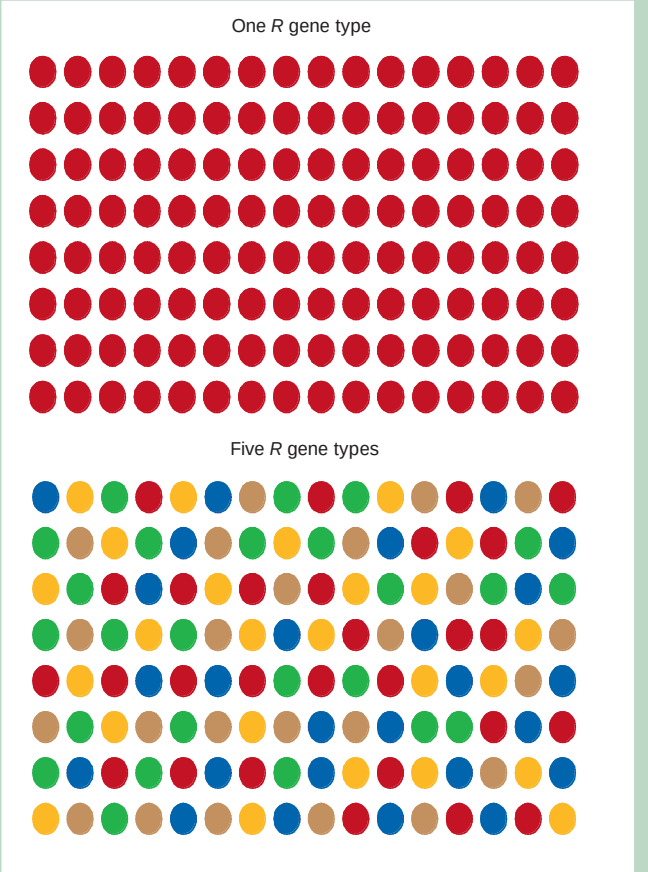


Figure 4 *R*-gene monoculture and *R*-gene polycultures. Different *R* genes are indicated by different colours. *R*-gene polycultures are proposed to give more durable resistance for three reasons. (1) Any pathogen race that can overcome only one *R* gene will give rise to a much slower epidemic. (2) Any such pathogen race that undergoes an additional mutation to overcome another *R* gene is likely to be less fit than a race that can overcome only one *R* gene, because *avr* genes are likely to encode pathogenicity factors. (3) High inoculum of avirulent races is likely to promote systemic acquired resistance, reducing the susceptibility of the plants to otherwise compatible pathogens.

defence pathway. Prf is thus an NB-LRR protein that 'guards' Pto, detects its interdiction by AvrPto (or any other bacterial effector), and then activates defence. More generally, in this model, R proteins physically associate with cellular targets of bacterial type III effectors of disease. These targets could include plant defence components or host proteins whose function is modified to nourish the extracellular bacterial colony. Generally, when the type III effector enters a resistant host cell, and interacts with a target, the resulting complex is recognized by the R protein, which is thus activated to initiate disease resistance. In the absence of a specific R protein, the host target is not guarded from the virulence function of the type III effector, and disease ensues.

In one mechanistic scenario (Fig. 5) the R protein could bind its guardee constitutively, but disengage upon type III effector binding to the guardee, resulting in an active R protein. This would be consistent with the observation that overexpression of *Prf* leads to constitutive disease resistance[73]. This model suggests that NB-LRR proteins, and the signalling pathways they mediate, are negatively regulated by their guardees. Alternatively, R-protein recruitment could be conditioned by engagement of the type III effector with its cellular target protein. A subsequent conformational change would then lead to enhanced affinity of the guardee–effector complex for the R protein, triggering resistance. Both models are consistent with the general lack of direct interaction between NB-LRR proteins and Avr proteins. Each scenario is consistent with the conceptual framework that R proteins monitor whether a cellular protein is under attack from a pathogen effector protein.

There are several predictions from the guard hypothesis. First, R proteins may interact constitutively with their 'guardees'. Such interactors will be targets of virulence factors required for successful infection, and/or components of defence signalling[74]. These possibilities are not mutually exclusive if the nominal target of virulence has also evolved a role in the stability of an R-containing complex whose structural integrity is required for resistance. For example, the viral coat protein of turnip crinkle virus is the Avr protein detected by *HRT* in this case. It interacts in yeast with an *Arabidopsis* NAC protein, probably a transcriptional regulator. Mutants in the coat protein that do not interact with this NAC protein are still virulent, but lose the ability to trigger *HRT*-mediated resistance[75]. Thus, this particular NAC protein is required for HRT function. It could also be required for basal defence and be a target for the coat protein, if the latter can also act as a virulence factor. Second, the complex of the effector with its target might be present in both susceptible and resistant host cells, except that in the latter the R protein will also be part of the complex. This suggests that a mutation in an R-protein partner might result in a loss of resistance conferred by that R protein, perhaps explaining the *pbs1* mutant phenotype. Third, if a limited number of host protein complexes are targets of the effector set from a given pathogen, then one host protein complex might be a target for multiple pathogen effector molecules. This suggests that multiple R proteins could physically associate with the same host protein complex and hence with each other. This notion may explain the functional interference of two structurally related R proteins[76,77], and the counter-intuitive finding that a particular type III effector could be co-immunoprecipitated with the 'wrong' R protein[78]. Fourth, a corollary of these predictions is that the number of host cellular targets of type III effectors may in fact be limited, and potentially targeted by several effectors. This is supported by at least two examples of a single R protein having the ability to recognize two different effectors[26,27]. In addition, the *Arabidopsis PBS1* gene is also a Ser/Thr kinase and could be the RPS5 'guardee'. Further support is provided by the fact that 46 of the *Arabidopsis* NB-LRR genes are single copy. This implies that they are ancient and provide effective resistance. Their effectiveness could be due to strict structural constraints on the corresponding type III effectors that are in turn targeted to a limited set of cellular targets. For the Cf-X class of R protein, Cf-9 could 'guard' the Avr9-binding site that is present even in stocks that lack the *Cf-9* gene[79].

The guard hypothesis is by no means proven, but it is does provide a step beyond the previous notion that R proteins are simply direct receptors for Avr proteins. This elicitor/receptor model may still be true for some systems, but for many others, the lack of direct R/Avr interaction is sufficiently convincing that it can be excluded.

Signal transduction and the effector arm

Genetic screens, almost exclusively in *Arabidopsis*, have defined loci required for *R*-gene action. Such loci are likely to encode proteins that function either as guardees (described above) or to mediate the series of biochemical events outlined below[2,80]. Note that a protein required for assembly of a cellular component targeted by a type III effector and guarded by an R protein might have multiple functions. Several mutants were identified by loss of a particular *R* function, and then tested for loss of additional *R* functions. Some of the resulting mutations are *R* specific, as they eliminate one specificity, whereas others define common steps in signal-transduction pathways required for the action of several *R* genes. This is a critical experimental step, as it is often easier to measure subtle effects on *R*-gene functions that are different than the one used in the original screening. These screens are inefficient; typically ~90% of the mutants are *r* alleles[81]. These results suggest that most mutations in the other genes required for the *R* signal pathway in question might be lethal, or that these steps are encoded by genes with overlapping or redundant functions.

© 2001 Macmillan Magazines Ltd

Loci required for basal defence have also been defined[2,80]. Mutants in these loci express enhanced disease susceptibility (*eds*) phenotypes and support more growth of virulent pathogens than the wild-type plant. Some of them also are required for the function of one or more *R* genes, and some are also required for pathogen-nonspecific systemic acquired resistance (SAR). Their requirement during *R*-dependent signalling proves that basal and specific defence systems can overlap.

Several genes required for multiple *R* functions are known. The *ndr1* and *eds1* mutants were defined in screens for loss of race-specific resistance to strains of the bacterium *P. syringae* or the oomycete *Peronospora parasitica*[82,83]. *EDS1* and *NDR1* are each required for the function of different NB-LRR *R* genes[84]. The *R* genes suppressed by the *ndr1* mutation are not affected by *eds1* mutants, and vice versa. *eds1* suppresses TIR-NB-LRR *R* genes, whereas *ndr1* suppresses a subset of CC-NB-LRR resistance proteins. Although these observations suggest a model in which *EDS1* and *NDR1* mediate distinct *R* gene-dependent signalling pathways, there are several examples of CC-NB-LRR *R* genes which function independently of both *EDS1* and *NDR1* (refs 84, 85). *RAR1* is required for the function of some, but not all, barley *Mla* resistance genes[86]. Differences in the amino-acid sequence between proteins encoded by *RAR1*-dependent and *RAR1*-independent *Mla* alleles are only around 5% (refs 87, 88). Thus, signalling proteins can discriminate between highly related R proteins. RAR1 is a novel zinc-coordinating protein, which, in higher metazoans, is co-linear with a protein domain homologous to the yeast SGT1 protein. SGT1 regulates delivery of targets to the SCF protein degradation machine[89]. Thus, upon activation, RAR1 may target a negative regulator of disease resistance and the hypersensitive response for degradation. Alternatively, the RAR1-containing SCF complex is a critical host target for a variety of powdery mildew effectors, and is therefore guarded by the products of several *Mla* alleles.

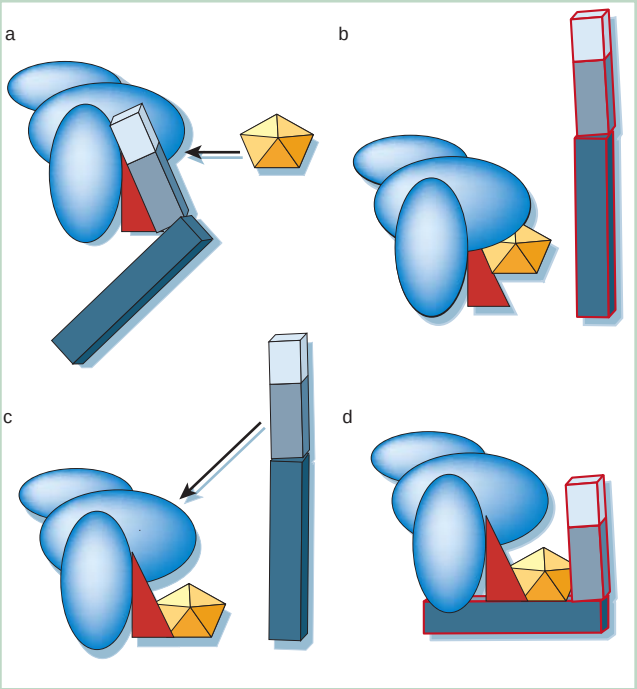


Figure 5 The guard hypothesis for R-protein function. **a**, A cellular complex of proteins (blue), which includes both the 'guardee' molecule (red) and an NB-LRR protein (grey, shaded from the N terminus through NB and LRR domains), is a target for a bacterial type III effector of disease (orange). **b**, Binding of the type III effector to its targets results in disassociation and activation of the NB-LRR protein and thus disease resistance. **c**, Alternatively, the NB-LRR protein may not be part of the target complex until after type III effector binding. **d**, Recruitment to the type III effector/target complex would then activate the NB-LRR protein.

The earliest events following *R* engagement are calcium influx, alkalinization of the extracellular space, protein kinase activation, production of reactive oxygen intermediates (ROIs) and nitric oxide, and transcriptional reprogramming. Studies using both nonspecific and race-specific elicitors have documented the chain of events that most rapidly ensue upon pathogen perception by plants. Plant cell cultures are more amenable to reproducible biochemical and pharmacological study than are whole plants. Elicitation of parsley cells with the PEP13 peptide derived from a non-race-specific elicitor from *Phytophthora megasperma*, elicitation of *Arabidopsis* cells with a conserved peptide derived from bacterial flagellin, and tobacco cells expressing the tomato *Cf-9* gene elicited by the Avr9 peptide system have revealed essentially similar processes[90–92]. Changes in ion fluxes, including calcium influx, occur within minutes of elicitation. Subsequently, ROIs (including H_2O_2 and/or O_2^-) are produced and mitogen-activated protein kinase and other protein kinase pathways are activated[93,94]. The ROIs may be involved in pathogen elimination, subsequent signalling of downstream effector functions, or (most likely) both. Studies on whole plants using bacterial strains recognized by NB-LRR R genes have revealed similar processes[95]. In addition, nitric oxide (NO) has been shown to accumulate through an as yet unidentified biosynthetic pathway[96–98].

Within 15 minutes, a set of new transcripts can be identified, comprising ~1% of total messenger RNA; these encode additional signalling molecules including protein kinases and transcription factors[99]. The protein kinases are upstream or independent of the oxidative burst, and presumably result in the activation of latent transcription factors required for defence gene activation[100]. NO and ROI could also contribute to rapid transcriptional activation of a battery of 'defence genes' in and surrounding the infected cell. Functions of these defence genes include biosynthesis of salicylic acid, induction of ethylene biosynthesis, cell-wall strengthening, lignification, production of various antimicrobial compounds, and a form of rapid cell death termed the 'hypersensitive response' (reviewed in ref. 101, and see review in this issue by Lam, Kato and Lawton, pages 848–853). It is, however, still unclear which of these events are causal mediators of *R*-gene action, and which are not. In addition to local resistance to infection, this set of events can also lead to establishment of SAR[102].

Transcriptional reprogramming establishes the 'effector' arm of the plant innate immune system. The defence response is clearly controlled by interacting signalling pathways. For example, in tobacco and *Arabidopsis*, enzymatic blocking of salicylic acid accumulation[103], or mutation of the *EDS16/SID2* which blocks salicylic acid biosynthesis[104], seriously impairs basal defence locally and induction of SAR in distal tissues. A key to understanding systemic signalling was the identification and isolation of the *Arabidopsis NPR1/NIM1* gene[105,106]. This gene transduces a salicylic acid-dependent signal to distal tissues, functions in a local basal-defence pathway, and is required for the action of a small number of tested *R* genes, but is not generally required for *R*-gene function. Ethylene signalling is also important in basal defence and can be required combinatorially with *NPR1/NIM1* for function of at least the *Arabidopsis* CC-NB-LRR gene *RPS2* (ref. 107). Jasmonic acid mediates wound responses and responses to necrotrophic fungal pathogens. Interestingly, the jasmonic acid signal pathway can act antagonistically to the salicylic acid pathway, allowing the response to be marshalled in a focused manner[108,109]. Large-scale studies of gene expression profiling are beginning to dissect this transcriptional output. The impact of co-regulatory circuits is beginning to be appreciated, with the finding that the *cis*-element bound by the plant-specific WRKY transcription factors is the common element in a set of SAR co-regulated genes[110].

The road at 'Rest and Be Thankful'

It is traditional in summing up to joyfully celebrate the past decade's substantial achievements by those on whose shoulders we now stand, while soberly and seriously looking ahead at the new horizons that have come into view. There is a parking area and scenic view on a

© 2001 Macmillan Magazines Ltd

small backroad in western Scotland between Loch Lomond and the sea where the tourist is bid to 'Rest and be Thankful'. Looking backwards, down the tortuous route climbed to this point, the traveller breathes relief. However, a look forward, down the road yet to run, reveals more of the same twists and drops. The field retains many enduring challenges and mysteries. Molecular geneticists need to grapple with identifying the *avr* genes of fungal pathogens that potentially traffic disease effectors into the host via haustoria. The completion of the *Arabidopsis* genome sequence, and the imminent completion of the genomes of several model bacterial plant pathogens (*Xylella fastidiosa* and *X. citri*, *Ralstonia solanacearum* and *P. syringae* DC3000) provide a rich new field for bioinformaticists and cell biologists to investigate gene function. It is to be hoped that in the next few years the genome sequences of rice blast (*M. grisea*), powdery mildew (*Blumeria graminis*) and other fungal pathogens will become available to the public sector. The availability of expression arrays and new proteomics tools will enable a complete transcriptional and post-transcriptional description of the defence response, at least in model plants. Such descriptive work is an essential prelude to further investigations of mechanisms. Despite the 7 years that have elapsed since the isolation of the first *R* genes, there is a great deal still to learn about how R proteins function to confer Avr recognition. The challenge of determining how R proteins work requires some change in focus towards biochemistry and cell biology. And despite some plausible interpretations, there is still a great need to do more field experiments to study how *R* genes work in natural populations, and to test approaches using genetic polymorphism to provide more durable disease resistance in crops. □

1. Flor, H. H. Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* **9,** 275–296 (1971).
2. Glazebrook, J., Rogers, E. E. & Ausubel, F. M. Use of Arabidopsis for genetic dissection of plant defense responses. *Annu. Rev. Genet.* **31,** 547–569 (1997).
3. Yang, Y., Shah, J. & Klessig, D. F. Signal perception and transduction in plant defense responses. *Genes Dev.* **11,** 1621–1639 (1997).
4. McDowell, J. M. & Dangl, J. L. Signal transduction in the plant innate immune response. *Trends Biochem. Sci.* **25,** 79–82 (2000).
5. Bent, A. Function meets structure in the study of plant disease resistance genes. *Plant Cell* **8,** 1757–1771 (1996).
6. Ellis, J., Dodds, P. & Pryor, T. Structure, function, and evolution of plant disease resistance genes. *Curr. Opin. Plant Biol.* **3,** 278–284 (2000).
7. Jones, J. D. G. Putting knowledge of plant disease resistance genes to work. *Curr. Opin. Plant Biol.* (in the press).
8. Boyes, D. C., Nam, J. & Dangl, J. L. The *Arabidopsis thaliana RPM1* disease resistance gene product is a peripheral plasma membrane protein that is degraded coincident with the hypersensitive response. *Proc. Natl Acad. Sci. USA* **95,** 15849–15854 (1998).
9. Jones, D. A. & Jones, J. D. G. The roles of leucine rich repeats in plant defences. *Adv. Bot. Res. Adv. Plant Pathol.* **24,** 90–167 (1996).
10. Kajava, A. V. Structural diversity of leucine-rich repeat proteins. *J. Mol. Biol.* **277,** 519–527 (1998).
11. Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biotechnol.* **15,** 430–435 (1990).
12. Aravind, L., Dixit, V. M. & Koonin, E. V. The domains of death: evolution of the apoptosis machinery. *Trends Biochem.* **24,** 47–53 (1999).
13. van der Biezen, E. A. & Jones, J. D. G. Homologies between plant resistance gene products and regulators of cell death in animals. *Current Biol.* **8,** R226–R227 (1998).
14. Parniske, M. *et al.* Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell* **91,** 821–832 (1997).
15. McDowell, J. M. *et al.* Intragenic recombination and diversifying selection contribute to the evolution of Downy Mildew resistance at the *RPP8* locus of Arabidopsis. *Plant Cell* **10,** 1861–1874 (1998).
16. Botella, M. A. *et al.* Three genes of the Arabidopsis *RPP1* complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell* **10,** 1847–1860 (1998).
17. Meyers, B. C. *et al.* The major resistance gene cluster in lettuce is highly duplicated and spans several megabases. *Plant Cell* **10,** 1817–1832 (1998).
18. Michelmore, R. W. & Meyers, B. C. Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res.* **8,** 1113–1130 (1998).
19. Wang, G.-L. *et al. Xa21D* encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. *Plant Cell* **10,** 765–779 (1998).
20. Ellis, J. G., Lawrence, G. J., Luck, J. E. & Dodds, P. N. Identification of regions in alleles of the flax rust resistance gene *L* that determine differences in gene-for-gene specificity. *Plant Cell* **11,** 495–506 (1999).
21. Wulff, B. B. H., Thomas, C. M., Smoker, M., Grant, M. & Jones, J. D. G. Domain swapping and gene shuffling identify specificity determinants required for induction of an Avr-dependent hypersensitive response by the tomato Cf-4 and Cf-9 proteins. *Plant Cell* **13,** 255–272 (2001).
22. Van der Hoorn, R. A. L., Roth, R. & De Wit, P. J. G. M. Identification of distinct specificity determinants in resistance protein Cf-4 allows construction of a Cf-9 mutant that confers recognition of avirulence protein AVR4. *Plant Cell* **13,** 273–285 (2001).
23. Luck, J. E., Lawrence, G. J., Dodds, P. N., Shepherd, K. W. & Ellis, J. G. Regions outside of the leucine-rich repeats of Flax Rust resistance proteins play a role in specificity determination. *Plant Cell* **12,** 1367–1377 (2000).
24. Initiative, T. A. G. Analysis of the genome of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
25. Bisgrove, S. R., Simonich, M. T., Smith, N. M., Sattler, N. M. & Innes, R. W. A disease resistance gene in Arabidopsis with specificity for two different pathogen avirulence genes. *Plant Cell* **6,** 927–933 (1994).
26. Grant, M. R. *et al.* Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. *Science* **269,** 843–846 (1995).
27. Rossi, M. *et al.* The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. *Proc. Natl Acad. Sci. USA* **95,** 9570–9754 (1998).
28. Cooley, M. B., Pathirana, S., Wu, H.-J., Kachroo, P. & Klessig, D. F. Members of the Arabidopsis *HRT/RPP8* family of resistance genes confer resistance to both viral and oomycete pathogens. *Plant Cell* **12,** 663–676 (2000).
29. van der Vossen, E. A. *et al.* Homologues of a single resistance-gene cluster in potato confer resistance to distinct pathogens: a virus and a nematode. *Plant J.* **23,** 567–576 (2000).
30. Stahl, E. A., Dwyer, G., Mauricio, R., Kreitman, M. & Bergelson, J. Dynamics of disease resistance polymorphism at the *RPM1* locus of *Arabidopsis*. *Nature* **400,** 667–671 (1999).
31. Riely, B. K. & Martin, G. B. Ancient origin of pathogen recognition specificity conferred by the tomato disease resistance gene Pto. *Proc. Natl Acad. Sci. USA* **98,** 2059–2064 (2001).
32. Dangl, J. L. in *Bacterial Pathogenesis of Plants and Animals: Molecular and Cellular Mechanisms* (ed. Dangl, J. L.) 99–118 (Springer, Heidelberg, 1994).
33. Medzhitov, R. *et al.* MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. *Mol. Cell* **2,** 253–258 (1998).
34. Dinesh-Kumar, S. P. & Baker, B. J. Alternatively spliced N resistance gene transcripts: their possible role in tobacco mosaic virus resistance. *Proc. Natl Acad. Sci. USA* **97,** 1908–1913 (2000).
35. Eulgem, T., Rushton, P. J., Robatzek, S. & Somssich, I. E. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* **5,** 199–206 (2000).
36. Grant, M. R. *et al.* Independent deletions of a pathogen-resistance gene in *Brassica* and *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **95,** 15843–15848 (1998).
37. Xiao, S. *et al.* Broad-spectrum mildew resistance in *Arabidopsis thaliana* mediated by *RPW8*. *Science* **291,** 118–120 (2001).
38. Noël, L. *et al.* Pronounced intraspecific haplotype divergence at the *RPP8* complex disease resistance locus of Arabidopsis. *Plant Cell* **11,** 2099–2111 (1999).
39. Scofield, S. R. *et al.* Molecular basis of gene-for-gene specificity in bacterial speck disease of tomato. *Science* **274,** 2063–2065 (1996).
40. Tang, X. *et al.* Physical interaction of avrPto and the Pto kinase defines a recognition event involved in plant disease resistance. *Science* **274,** 2060–2063 (1996).
41. Salmeron, J. M. *et al.* Tomato *Prf* is a member of the leucine-rich repeat class of plant disease resistance genes and lies embedded within the *Pto* kinase gene cluster. *Cell* **86,** 123–133 (1996).
42. Song, W.-Y. *et al.* A receptor kinase-like protein encoded by the rice disease resistance gene, *Xa21*. *Science* **270,** 1804–1806 (1995).
43. He, Z. *et al.* Perception of Brassinosteroids by the extracellular domain of the receptor kinase BRI1. *Science* **288,** 2360–2363 (2000).
44. Wang, Z.-Y., Seto, H., Fujioka, S., Yoshida, S. & Chory, J. BRI1 is a critical component of a plasma-membrane receptor for plant steroids. *Nature* **410,** 380–383 (2001).
45. Jones, D. A., Thomas, C. M., Hammond-Kosack, K. E., Balint-Kurti, P. J. & Jones, J. D. G. Isolation of the tomato *Cf-9* gene for resistance to *Cladosporium fulvum* by transposon tagging. *Science* **266,** 789–793 (1994).
46. Trotochaud, A. E., Hao, T., Wu, G., Yang, Z. & Clark, S. E. The *CLAVATA1* receptor-like kinase requires *CLAVATA3* for its assembly into a signaling complex that includes KAPP and a Rho-related protein. *Plant Cell* **11,** 393–405 (1999).
47. Gómez-Gómez, L. & Boller, T. FLS2: an LRR receptor like kinase involved in the perception of the bacterial elicitor flagellin in Arabidopsis. *Mol. Cell* **5,** 1003–1011 (2000).
48. Swiderski, M. R. & Innes, R. W. The Arabidopsis *PBS1* resistance gene encodes a member of a novel protein kinase subfamily. *Plant J.* **26,** 101–112 (2001).
49. Aderem, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406,** 782–787 (2000).
50. Medzhitov, R. & Janeway, C. Jr The Toll receptor family and microbial recognition. *Trends Microbiol.* **8,** 452–456 (2000).
51. Ozinsky, A. *et al.* The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between Toll-like receptors. *Proc. Natl Acad. Sci. USA* **97,** 13766–13771 (2000).
52. Inohara, N., Ogura, Y., Chen, F. & Nunez, G. Human Nod1 confers responsiveness to bacterial lipopolysaccharides. *J. Biol. Chem.* **276,** 2551–2554 (2001).
53. Ogura, Y. *et al.* Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF-κB. *J. Biol. Chem.* **276,** 4812–4818 (2001).
54. Hugot, J.-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411,** 599–603 (2001).
55. Ogura, Y. *et al.* A frameshift in *Nod2* associated with susceptibility to Crohn's disease. *Nature* **411,** 603–606 (2001).
56. Hartman, J. L. IV, Garvik, B. & Hartwell, L. Principles for the buffering of genetic variation. *Science* **291,** 1001–1004 (2001).
57. Hamilton, W. D., Axelrod, R. & Tanese, R. Sexual reproduction as an adaptation to resist parasites. *Proc. Natl Acad. Sci. USA* **87,** 3566–3573 (1990).
58. Gálan, J. E. & Collmer, A. Type III secretion machines: bacterial devices for protein delivery into host cells. *Science* **284,** 1322–1328 (1999).
59. Hueck, C. J. Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol. Mol. Biol. Rev.* **62,** 379–433 (1998).
60. Kjemtrup, S., Nimchuk, Z. & Dangl, J. L. Effector proteins of phytopathogenic bacteria: bifunctional signals in virulence and host recognition. *Curr. Opin. Microbiol.* **3,** 73–78 (2000).
61. Gopalan, S. *et al.* Expression of the *Pseudomonas syringae* avirulence protein AvrB in plant cells alleviates its dependence on the hypersensitive response and pathogenicity (Hrp) secretion system in eliciting genotype-specific hypersensitive cell death. *Plant Cell* **8,** 1095–1105 (1996).
62. McNellis, T. W. *et al.* Glucocorticoid-inducible expression of a bacterial avirulence gene in transgenic *Arabidopsis thaliana* induces hypersensitive cell death. *Plant J.* **14,** 247–258 (1998).
63. Nimchuk, Z. *et al.* Eukaryotic fatty acylation drives plasma membrane targeting and enhances function of several Type III effector proteins from *Pseudomonas syringae*. *Cell* **101,** 353–363 (2000).
64. Mudgett, M. & Staskawicz, B. Characterization of the Pseudomonas syringae pv. tomato AvrRpt2

© 2001 Macmillan Magazines Ltd

protein: demonstration of secretion and processing during bacterial pathogenesis. *Mol. Microbiol.* **32,** 927–941 (1999).
65. Nimchuk, Z., Rohmer, L., Chang, J. H. & Dangl, J. L. Knowing the dancer from the dance: R gene products and their interactions with other proteins from host and pathogen. *Curr. Opin. Plant Biol.* (in the press).
66. Laugé, R., Joosten, M. H. A., Van den Ackerveken, G. F. J. M., Van den Broek, H. W. J. & De Wit, P. J. G. M. The in-planta produced extracellular proteins ECP1 and ECP2 of *Cladosporium fulvum* are virulence factors. *Mol. Plant-Microbe Interact.* **10,** 725–734 (1997).
67. Laugé, R. *et al.* Successful search for a resistance gene in tomato targeted against a virulence factor of a fungal pathogen. *Proc. Natl Acad. Sci. USA* **95,** 9014–9018 (1998).
68. Laugé, R., Dmitriev, A. P., Joosten, M. H. A. J. & De Wit, P. G. J. M. Additional resistance gene(s) against *Cladosporium fulvum* present on the *Cf-9* introgression segment are associated with strong PR protein accumulation. *Mol. Plant-Microbe Interact.* **11,** 301–308 (1998).
69. Lauge, R., Goodwin, P. H., de Wit, P. J. & Joosten, M. H. Specific HR-associated recognition of secreted proteins from Cladosporium fulvum occurs in both host and non-host plants. *Plant J.* **23,** 735–745 (2000).
70. Kooman-Gersman, M., Honée, G., Bonnema, G. & De Wit, P. J. G. M. A high-affinity binding site for the AVR9 peptide elicitor of *Cladosporium fulvum* is present on plasma membranes of tomato and other Solanaceous plants. *Plant Cell* **8,** 929–938 (1996).
71. Bryan, G. T. *et al.* A single amino acid difference distinguishes resistant and susceptible alleles of the rice resistance gene *Pi-ta*. *Plant Cell* **12,** 2033–2045 (2000).
72. van der Biezen, E. A. & Jones, J. D. G. Plant disease resistance proteins and the "gene-for-gene" concept. *Trends Biochem. Sci.* **23,** 454–456 (1998).
73. Oldroyd, G. E. D. & Staskawicz, B. J. Genetically engineered broad-spectrum disease resistance in tomato. *Proc. Natl Acad. Sci. USA* **95,** 10300–10305 (1998).
74. Chen, Z., Kloek, A. P., Boch, J., Katagiri, F. & Kunkel, B. N. The *pseudomonas syringae avrRpt2* gene product promotes pathogenicity from inside the plant cell. *Mol. Plant Pathol.* **13,** 1312–1321 (2000).
75. Ren, T., Qu, F. & Morris, T. J. *HRT* gene function requires interaction between a NAC protein and viral capsid protein to confer resistance to Turnip Crinkle Virus. *Plant Cell* **12,** 1917–1925 (2000).
76. Ritter, C. & Dangl, J. L. Interference between two specific pathogen recognition events mediated by distinct plant disease resistance genes. *Plant Cell* **8,** 251–257 (1996).
77. Reuber, T. L. *et al.* Correlation of defense gene induction defects with powdery mildew susceptibility in *Arabidopsis* enhanced disease susceptibility mutants. *Plant J.* **16,** 473–485 (1998).
78. Leister, R. T. & Katagiri, F. A resistance gene product of the nucleotide binding site-leucine rich repeats class can form a complex with bacterial avirulence proteins *in vitro*. *Plant J.* **22,** 345–354 (2000).
79. Dixon, M. S., Golstein, C., Thomas, C. M., van Der Biezen, E. A. & Jones, J. D. Genetic complexity of pathogen perception by plants: the example of Rcr3, a tomato gene required specifically by Cf-2. *Proc. Natl Acad. Sci. USA* **97,** 8807–8814 (2000).
80. Glazebrook, J. Genes controlling expression of defense responses in Arabidopsis—2001 status. *Curr. Opin. Plant Biol.* (in the press).
81. Jorgensen, J. H. Genetic analysis of barley mutants with modifications of the powdery mildew resistance gene Mla_{12}. *Genome* **30,** 129–132 (1988).
82. Century, K. S., Holub, E. B. & Staskawicz, B. J. *NDR1*, a locus of *Arabidopsis thaliana* that is required for disease resistance to both a bacterial and a fungal pathogen. *Proc. Natl Acad. Sci. USA* **92,** 6597–6601 (1995).
83. Parker, J. E. *et al.* Characterization of *eds1*, a mutation in Arabidopsis suppressing resistance to *Peronospora parasitica* specified by several different *RPP* genes. *Plant Cell* **8,** 2033–2046 (1996).
84. Aarts, N. *et al.* Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene mediated signalling pathways in Arabidopsis. *Proc. Natl Acad. Sci. USA* **95,** 10306–10311 (1998).
85. McDowell, J. M. *et al.* Downy mildew (*Peronospora parasitica*) resistance genes in Arabidopsis vary in functional requirements for *NDR1*, *EDS1*, *NPR1*, and Salicylic Acid accumulation. *Plant J.* **22,** 523–530 (2000).
86. Freialdenhoven, A. *et al. Nar-1* and *Nar-2*, two loci required for *Mla-12*-specified race resistance to powdery mildew in barley. *Plant Cell* **6,** 983–994 (1994).
87. Zhou, F. *et al.* Cell-autonomous expression of barley *Mla1*confers race specific resistance to the powdery mildew fungus via a *Rar1* independent signaling pathway. *Plant Cell* **13,** 337–350 (2001).
88. Halterman, D., Zhou, F., Wei, F., Wise, R. P. & Schulze-Lefert, P. The Mla6 coiled-coil, NBS-LRR protein functions in barley and wheat to confer resistance specificity to *Blumeria graminis* f. sp. *hordei*. *Plant J.* (in the press).
89. Kitigawa, K., Skowyra, D., Elledge, S. J., Harper, J. W. & Hieter, P. *SGT1* encodes an essential component of the yeast kinetochore assembly pathway and a novel subunit of the SCF ubiquitin complex. *Mol. Cell* **4,** 21–33 (1999).
90. Jabs, T., Colling, C., Tschöpe, M., Hahlbrock, K. & Scheel, D. Elicitor-stimulated ion fluxes and reactive oxygen species from the oxidative burst signal defense gene activation and phytoalexin synthesis in parsley. *Proc. Natl Acad. Sci. USA* **94,** 4800–4805 (1997).
91. Felix, G., Duran, J. D., Volko, S. & Boller, T. Plants have a sensitive perception system for the most conserved domain of bacterial flagellin. *Plant J.* **18,** 265–276 (1999).
92. Piedras, P., Hammond-Kosack, K. E., Harrison, K. & Jones, J. D. G. Rapid, Cf-9 and Avr9 dependent, production of active oxygen species in tobacco suspension cultures. *Mol. Plant-Microbe Interact.* **11,** 1155–1166 (1998).
93. Ligternik, W., Kroj, T., zur Nieden, U., Hirt, H. & Scheel, D. Receptor-mediated activation of a MAP kinase in pathogen defense of plants. *Science* **276,** 2054–2057 (1997).
94. Romeis, T. *et al.* Rapid, Avr9- and Cf-9-dependent activation of MAP kinases in tobacco cell cultures and leaves: convergence of resistance gene, elicitor, wound and salicylate responses. *Plant Cell* **11,** 273–287 (1999).
95. Grant, M. *et al.* The *RPM1* plant disease resistance gene facilitates a rapid and sustained increase in cytosolic calcium that is necessary for the oxidative burst and hypersensitive cell death. *Plant J.* **23,** 441–450 (2000).
96. Durner, J., Wendehenne, D. & Klessig, D. F. Defense gene induction in tobacco by nitric oxide, cyclic GMP and cyclic ADP ribose. *Proc. Natl Acad. Sci. USA* **95,** 10328–10333 (1998).
97. Delledonne, M., Xia, Y., Dixon, R. A. & Lamb, C. J. Nitric oxide functions as a signal in plant disease resistance. *Nature* **394,** 585–588 (1998).
98. Delledonne, M., Zeier, J., Marocco, A. & Lamb, C. J. Signal interactions between nitric oxide and reactive oxygen intermediates in the plant hypersensitive disease resistance response. *Proc. Natl Acad. Sci. USA* (in the press).
99. Durrant, W. E., Rowland, O., Piedras, P., Hammond-Kossak, K. E. & Jones, J. D. G. cDNA-AFLP reveals a striking overlap in the race-specific resistance and wound response expression profiles. *Plant Cell* **12,** 963–977 (2000).
100. Dröge-Laser, W. *et al.* Rapid stimulation of a soybean protein serine-threonine kinase which phosphorylates a novel bZIP DNA-binding protein, G/HBF-1, during the induction of early transcription-dependent defenses. *EMBO J.* **16,** 726–738 (1997).
101. Scheel, D. Resistance response physiology and signal transduction. *Curr. Opin. Plant Biol.* **1,** 305–310 (1998).
102. Delaney, T. P. Genetic dissection of acquired resistance to disease. *Plant Physiol.* **113,** 5–12 (1997).
103. Gaffney, T. *et al.* Requirement for salicylic acid for the induction of systemic acquired resistance. *Science* **261,** 754–756 (1993).
104. Nawrath, C. & Métraux, J.-P. Salicylic acid induction-deficient mutants of Arabidopsis express PR-2 and PR-5 and accumulate high levels of camalexin after pathogen attack. *Plant Cell* **11,** 1393–1404 (1999).
105. Cao, H., Glazebrook, J., Clark, J. D., Volko, S. & Dong, X. The Arabidopsis *NPR1* gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell* **88,** 57–64 (1997).
106. Ryals, J. *et al.* The Arabidopsis *NIM1* protein shows homology to the mammalian transcription factor inhibitor IκB. *Plant Cell* **9,** 425–439 (1997).
107. Clarke, J. D., Volko, S. M., Ledford, H., Ausubel, F. M. & Dong, X. Roles of salicylic acid, jasmonic acid, and ethylene in *cpr*-induced resistance in Arabidopsis. *Plant Cell* **12,** 2175–2190 (2000).
108. Reymond, P. & Farmer, E. E. Jasmonate and salicylate as global signals for defense gene expression. *Curr. Opin. Plant Biol.* **1,** 404–411 (1998).
109. Reymond, P., Weber, H., Damond, M. & Farmer, E. E. Differential gene expression in response to mechanical wounding and insect feeding in Arabidopsis. *Plant Cell* **12,** 707–719 (2000).
110. Maleck, K. *et al.* The transcriptome of Arabidopsis during systemic acquired resistance. *Nature Genet.* **26,** 403–410 (2000).

Acknowledgements

Work in the Dangl laboratory is supported by grants from the NSF, NIH, DOE, USDA-NRI and Syngenta. The Jones laboratory is supported by the Gatsby Charitable Trust and the BBSRC. We thank P. Epple, T. Eulgem, J. McDowell, S. Peck, J. Rathjen and B. Staskawicz for critical reading of this manuscript, and G. Nuñez and D. Golenbock for useful suggestions.

© 2001 Macmillan Magazines Ltd

Gene silencing as an adaptive defence against viruses

Peter M. Waterhouse*, Ming-Bo Wang* & Tony Lough†

**CSIRO Plant Industry, Canberra, ACT 2601, Australia (e-mail: peterw@pi.csiro.au)*
†HortResearch, Tennent Drive, Palmerston North, New Zealand (Present address: Genesis Research and Development Corporation, PO Box 50, Auckland, New Zealand)

Gene silencing was perceived initially as an unpredictable and inconvenient side effect of introducing transgenes into plants. It now seems that it is the consequence of accidentally triggering the plant's adaptive defence mechanism against viruses and transposable elements. This recently discovered mechanism, although mechanistically different, has a number of parallels with the immune system of mammals.

Biology students are taught that the concept of vaccination came from Edward Jenner's discovery that milkmaids and dairymen infected with the mild cowpox virus were protected against smallpox. It is less widely appreciated that plants can also be protected from a severe virus by prior infection with a mild strain of a closely related virus. This cross protection in plants was recognized as early as the 1920s, but its mechanism has been a mystery — plants do not possess an antibody-based immune system analogous to that found in animals. This was probably the first observation of a plant's intrinsic defence mechanism against viruses which, 75 years later, is just beginning to be understood. In the past decade, there has been considerable research into transgene-mediated virus resistance, co-suppression, virus-induced gene silencing (VIGS), antisense suppression and transcriptional gene silencing (TGS) in plants. There has also been intense research into RNA interference in *Drosophila* and nematodes, and quelling in fungi. These seemingly disparate endeavours have produced pieces of a jigsaw puzzle which, when put together, begin to reveal the existence and characteristics of a natural defence system in plants against viruses and transposable DNA elements. Many of the details and ramifications have yet to be determined, but the current picture is that of a wonderfully elegant system that can generically recognize invading viruses and transposable elements (TEs) and marshal the plant's defences against them.

Plant viruses and transposable DNA elements

There are currently 72 different defined genera of plant viruses[1], containing over 500 species, and there is scarcely a plant species — mono- or dicotyledon — that is not host to at least one virus. Plant viruses have a whole array of different particle morphologies, host ranges, vectors (for example, insects, nematodes, fungi, pollen, seeds or humans), genome organizations and gene expression strategies. They cause symptoms which at their least severe are unnoticeable, but range upwards through ringspots or mosaic leaf patterns, to widespread necrosis. The genomes of some plant viruses are encoded using single-stranded (ss) or double-stranded (ds) DNA; others have dsRNA genomes. However, over 90% of plant viruses have ssRNA genomes that are replicated by a virus-encoded RNA-dependent RNA polymerase (RDRP). Plants defend themselves by exploiting this requirement of most plant viruses to replicate using a double-stranded replicative intermediate.

TEs are DNA sequences that have the capacity to move from place to place within a genome. They have been divided into two classes. Class I TEs are retroelements that amplify their copy number through reverse transcription of an RNA intermediate. They are particularly abundant in eukaryotes, and in plants comprise the greatest mass of TEs (in maize, this class of TE makes up over 70% of the nuclear DNA). Of the four types of retroelements in plants, the main class contains retrotransposons with direct long terminal repeats (LTRs). Class II TEs occur in all organisms, particularly prokaryotes; they have terminal inverted repeats (TIRs) ranging in size from 11 to several hundred base pairs. Within a class II TE family, one or more elements encode a transposase that has the potential to interact with TIRs to excise the elements and integrate them into other regions of the genome (for recent reviews of plant TEs, see refs 2, 3). Both classes of TEs can move around plant genomes, altering the function and structure of genes, and so accelerating genomic evolution. However, they are also parasitic mutagenic agents that have the potential to lacerate a genome[4]. To ensure survival, a plant needs to keep TE activity in check.

Targeted RNA degradation

Although not recognized at the time, evidence of a plant's intrinsic defence mechanism to counter viruses and transposons came from the initially mystifying results of co-suppression and transgene-mediated virus resistance. Transformation with antisense gene constructs has been used in plant research since 1987 (ref. 5). From previous work on natural antisense in bacteria[6], it was thought that hybridization of antisense RNA to the target messenger RNA interfered with its transport or translation[7]. So it was surprising to find subsequently that transformation of plants with transgene constructs encoding sense mRNA homologous to endogenous genes could also suppress the activities of these genes[8–10]. It was similarly perplexing when plants transformed with virus-derived transgenes, designed to provide protection through a protein-mediated mechanism[11], gave protection against viruses even when little or no transgene protein (transprotein) was produced[12] (Fig. 1).

Further analysis of the co-suppressed and virus-resistant transprotein-free plants revealed that in both cases

© 2001 Macmillan Magazines Ltd

the transgenes were being highly transcribed in the nucleus, but the steady-state levels of their mRNAs in the cytoplasm were very low. This led to the proposal[12,13] that the transgene mRNA was somehow perceived by the cell as unwanted and induced sequence-specific degradation, by a targeted nuclease, of itself and other homologous or complementary RNA sequences in the cytoplasm. Thus, in the co-suppressed and virus-resistant lines, not only the transgene mRNAs but also the mRNA from the homologous endogenous gene and the invading virus RNA (with homology to the transgene) were being degraded. The concept of transgene RNA-directed RNA degradation was supported by the results of an experiment in which plants, with a co-suppressed β-glucuronidase (GUS) reporter gene, were inoculated with a wild-type plant virus or the same virus engineered to contain GUS sequences in its genome. The plants were susceptible to the wild-type virus, but resistant to the virus containing the GUS-encoding sequence. The virus has an RNA genome and replicates exclusively in the cytoplasm, so the simple explanation is that the GUS sequence within the virus genome was specifically degraded in the cytoplasm by the same mechanism that was causing co-suppression of the nuclear-expressed genes[14].

This conclusion raised a number of questions. How do the nucleases in the cell know which RNAs to degrade and which to leave alone, or more specifically, how do they distinguish transgene RNA from endogenous gene mRNA? Why does this not happen to the mRNAs from all transgenes? And why would a plant want to specifically degrade these RNAs? A critical observation was that in both the co-suppression and virus-transgene transformation experiments, only a proportion of the initial transformants showed co-suppression or virus resistance, and these plants generally contained multiple, methylated copies of the transgenes.

How is the degradation system triggered?

Several theories have been advanced to explain how this sequence-specific degradation system might be activated. It was proposed initially that the high copy number of the transgenes produced excessively high levels of transgene mRNA and that this level induced the degradation system[12,13]. Other researchers suggested that the methylation of the transgenes made them produce aberrant (for example, prematurely terminated) RNA and that this aberrance induced the system[14–17]. A compelling proposal was that the system is induced and directed by dsRNA and that multiple transgenes favoured the likelihood of their integration as inverted repeats which, by transcriptional readthrough from one transgene into the other, would produce duplex-forming, self-complementary RNA[18]. This was supported by the demonstration that transgenes deliberately designed to produce self-complementary (hairpin or hp) RNA or dsRNA were highly efficient at inducing targeted virus resistance and gene silencing[18–20]. Furthermore, an investigation of simple co-suppression and antisense constructs found a perfect correlation between the integration of these constructs as inverted repeats and the induction of silencing[19], and analysis of similar loci detected the presence of hpRNAs transcribed from them[21].

Why would the plant want to degrade dsRNA and ssRNAs of similar sequence? Healthy plants do not contain dsRNA or extensively self-complementary ssRNA. In fact, for many years, plant virologists have used the presence of dsRNA in plant extracts to diagnose viral infection[22]. This seems to be the key. Most plant viruses have ssRNA genomes and replicate in the cytoplasm using their own RDRP to produce both sense and antisense (termed plus-strand and minus-strand) RNA. Evidence from research on the RNA bacteriophage Qβ[23] suggests that the plus and minus strands of a ssRNA virus form full-length dsRNA only as an artefact of extraction[24]. However, when the mammalian 2′,5′-oligoadenylate system (which is activated specifically by dsRNA) was transformed into plants, it was activated by infection with either of the two ssRNA plant viruses tested[25]. Therefore, replication of a ssRNA plant virus can produce sufficient stretches of dsRNA to be recognized as such within the plant. This is consistent with the phenomenon of VIGS; here, plant viruses that contain sequences homologous to nuclear-expressed genes act to induce silencing of the targeted genes[26].

It therefore seems likely that one of the roles of the dsRNA-induced RNA degradation system of plants is to protect them against virus infection. This is a way to generically detect the replication of an invading ssRNA virus and destroy it, by specifically degrading both replicating and translatable forms of its genome.

Nuclease specificity and location

Specific fragments from mRNAs or viral genomes have been identified in gene-silenced or virus-resistant tissues, indicating that the targeted RNA degradation starts with endonucleolytic cleavage at one or more sites and is followed by exonucleolytic degradation[14,27,28]. Further investigation has found that sense and antisense ~25-nucleotide RNAs, with homology to the target RNA, are found consistently in plants showing co-suppression, antisense suppression, VIGS and virus resistance, but not in the appropriate control plants[29–33]. This is a critical finding. It supplies further evidence that these different forms of silencing are all acting by the same mechanism; from here on we refer to them generically as post-transcriptional gene silencing (PTGS). It also provides a strong link between PTGS and a phenomenon called RNA interference (RNAi), which is the targeted inhibition of gene activity by introduction, usually by injection, of dsRNA into a number of lower eukaryotes, including nematodes and *Drosophila*[34,35].

Looking at the biochemistry of the process of RNAi provides a good indication of what is probably happening in plants (Fig. 2). The processes of RNAi have been examined in *Drosophila* embryos, and embryo extracts, using radiolabelled dsRNA and target ssRNA[36–40]. Target ssRNA is not significantly degraded when sense or antisense RNAs are also introduced. However, the target RNA is degraded within minutes of adding homologous dsRNA. The degradation rapidly produces short sense and antisense ~21-nucleotide RNAs from both the dsRNA and the target ssRNA as a two-step process[39,40]. The dsRNA is degraded in ~21-nucleotide steps from both ends by an enzyme called Dicer-1 (CG64792, *DCR1*). The cleavage process, which is similar to that of *Escherichia coli* RNase III, produces ~21-nucleotide dsRNA fragments with 3′ overhangs of 2–3 nucleotides, and 5′-phosphate and 3′-hydoxyl termini[40]. Each fragment is associated with, and cleaved by, a separate Dicer-containing complex. The current model for the second step of the degradation is that the Dicer-containing, small interfering ribonucleoprotein (siRNP) complex alters in such a way that the strands of short dsRNA become unpaired and guide the complex to complementary target RNAs. This probably requires recruitment of

Figure 1 Potato plants challenged with potato virus Y. The three plants on the right are non-transgenic and are susceptible to the virus. The three plants on the left contain an untranslatable virus-derived transgene yet are immune to the virus[18].

© 2001 Macmillan Magazines Ltd

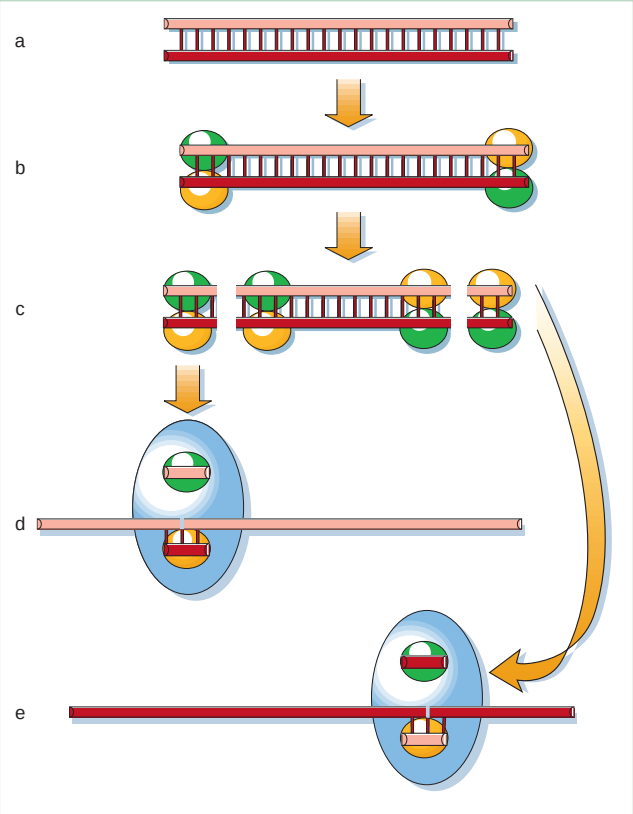


Figure 2 Proposed mechanism, based on RNAi, for dsRNA-directed ssRNA cleavage in PTGS. Introduced dsRNA (**a**) attracts Dicer-1-like proteins to its termini (**b**). The heterodimer complex at each end cleaves a 21-nucleotide dsRNA fragment (**c**), and the exposed ends of the shortened dsRNA each attract a new Dicer complex, which cleaves a further 21-nucleotide dsRNA fragment. This progressive exonuclease-like shortening continues until the dsRNA is completely cleaved. Dicers, loaded with dsRNA, acquire further components (blue ellipse), melt their dsRNA fragments and use one strand to hybridize to homologous ssRNA and cleave it in the middle of the 21-nucleotide guide-recognized sequence. One half of the dimer (shown as gold) directs hybridization and endonuclease cleavage giving it sense specificity. Thus Dicer complexes loaded from one end of a dsRNA will cleave sense-strand mRNA (**d**), whereas complexes loaded from the other end will cleave mRNA of the opposite sense (**e**).

additional proteins to the complex[39]. Once hybridized to a target RNA, the complex cleaves it at a position approximately in the middle of the recognized 21-nucleotide sequence. The whole two-step process results in dsRNA being cleaved with a ~21-nucleotide (two helical turns[36]) periodicity from their termini, and the appropriate target RNAs being cleaved with the same periodicity but with a frame shift of ~10 nucleotides.

The second step of the degradation probably takes place exclusively in the cytoplasm, as silencing does not reduce the full-length transcript levels in the nucleus[27]. However, the first step could occur in both the cytoplasm and the nucleus. Many mRNA degradation mechanisms involve the association of RNA with ribosomes, so it might be assumed that this would be the site of siRNP-mediated degradation. But several studies using protein-synthesis inhibitors have shown that neither ongoing translation nor association of the target RNAs with the ribosome are required for this degradation[15,41]. Furthermore, for each ribosome to be associated with enough siRNP complexes to ensure effective degradation of target RNA, the siRNPs would have to be expressed at very high levels. Perhaps the degradation complexes act as gatekeepers, located at the nuclear pores and plasmodesmata, scanning the RNAs that pass through. This would allow mRNAs to be efficiently screened and destroyed, if recognized, as they exit the nucleus, thus leading to gene silencing. Viral RNAs would be scanned and destroyed as they attempt to spread from cell to cell through the plasmodesmata.

Table 1 Genetic components of post-transcriptional silencing pathways

Gene function	Plants*	Worms*	Flies*	Fungi*	Algae*	TE act.†	PAZ‡
RDRP	SGS2 SDE1§	EGO1		QDE1			
Translation initiation factor	AGO1	RDE1		QDE2		No	Yes
RecQ, DEAH or Upf1p helicase	SDE3	MUT7‖ SMG-2		QDE3	MUT6‖	Yes‖	
RNaseIII + helicase¶	CAF1	K12H4.8	DCR1				Yes
Chromatin remodelling	DDM1					Yes	
Methyl transferase	MET1						
?	SGS3						
?		RDE 4				No	
?		RDE2&3 MUT2				Yes	
Methylation	Yes?	No	No	No			

*Plants, *Arabidopsis thaliana*; worms, *Caenorhabditis elegans*; flies, *Drosophila melanogaster*; fungi, *Neurospora crassa*; algae, *Chlamydomonas reinhardtii*.
†TE act., activation of transposon activity in organisms mutant for this gene function.
‡PAZ, presence of the PAZ domain identified by Cerruti *et al.*[54].
§SGS2 and SDE1 are different descriptions of the same gene.
‖The effect on transposon activation was assessed only for the MUT7 and MUT6 helicases. MUT7 also has an RNAse D motif.
¶CAF1 and K12H4.8 have been identified by sequence homology to DCR1 but they have not been shown formally to be involved in PTGS.

What genes are involved in PTGS and RNAi?

The similarity of induction, degradation and associated short dsRNAs in RNAi, quelling and PTGS indicates an underlying evolutionarily conserved mechanism. Analysis of mutants defective in these processes in *Caenorhabditis elegans*, *Neurospora* and *Arabidopsis* confirm this closeness, showing that there are a number of common essential enzymes or factors (Table 1). In all three species, mutation of an RDRP or a protein with homology to eIF2C, a rabbit protein thought to be involved in translation initiation[42], blocks silencing[32,43–48]. Another class of essential silencing proteins, those with homology to one of three types (RecQ, DEAH or Upf1p) of helicase, has been found in *C. elegans*, *Neurospora*, *Chlamydomonas reinhardtii* and *Arabidopsis*[49–53]. The roles of these proteins remain to be elucidated. They are probably not the equivalents (or parts thereof) of Dicer-1 in *Drosophila*; comparisons of the Dicer-1 sequence with genome databases identify K12H4.8 in *C. elegans* and CAF1 in *Arabidopsis* as homologues. These two Dicer-like proteins each have an RNA helicase domain, RNase III motifs and a PAZ domain[39,54] (Fig. 3).

There are two further categories of silencing-deficient mutants in plants and nematodes. One contains mutations of proteins that affect the structure and/or transcriptional status of chromatin, including DDM1, which remodels chromatin structure, MET1, which is a methyltransferase, and RDE2, RDE3 and MUT2, which seem to be involved with repressing the activity of TEs. The other category contains SGS3 in *Arabidopsis* and RDE4 from *C. elegans*, whose functions are a complete mystery. SGS3 has been cloned and sequenced, but has no recognizable motifs or matches with other sequences in available databases.

These mutation studies show that PTGS, RNAi and quelling are not just the result of Dicer complexes waiting to degrade dsRNA and homologous ssRNA that invades a cell. Other associated processes are clearly involved, including a possible link to the translation apparatus, an RDRP, and interactions with chromosomal DNA.

The role of methylation and chromatin remodelling in PTGS

DNA methylation and chromatin structure have an integral role in TGS. In this form of silencing, the promoter and sometimes the

© 2001 Macmillan Magazines Ltd

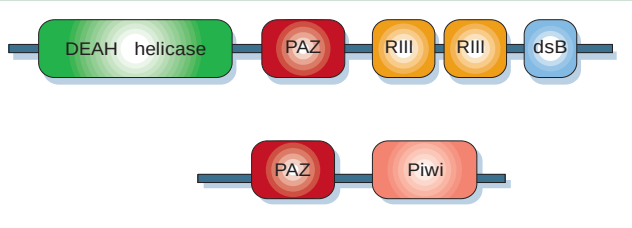


Figure 3 Distribution of domains on DCR1-like and AGO1-like proteins. Top panel illustrates the domains on DCR1-like proteins of ~2,000 amino acids (including DCR1, CAF1 and K12H4.8); bottom panel represents AGO1-like proteins of ~900 amino acids (including AGO1, RDE1 and QDE2). RIII, RNase III domain; dsB, double-stranded RNA-binding domain(s). For more detailed information on domains, see http://www.sanger.ac.uk/Software/Pfam.

coding region of the silenced transgenes are densely methylated[55]. Methylation, or methylation-associated chromatin remodelling, of promoter sequences is thought to prevent binding of factors necessary for transcription[55]. The coding sequences of PTGS-inducing transgenes are also frequently found to be methylated. PTGS can be established in plants with a mutant methyltransferase (*metI*), but during growth, the silencing becomes impaired, reactivating the silenced gene in sectors of the plant[56]. Furthermore, PTGS can fail to establish in mutant plants lacking the chromatin remodelling protein DDM1. These results suggest a role for DNA methylation and/or chromatin structure in both establishment and maintenance of PTGS.

The mechanisms of PTGS and TGS may have more in common than was previously thought. In PTGS, the short RNAs derived from the transcribed region of the transgene act as guides for siRNPs to degrade target ssRNA. In TGS plants, hpRNAs containing promoter-region sequences are processed into short dsRNAs, and seem to direct methylation[30]. Similarly, virus-replicated RNAs direct sequence-specific DNA methylation[57,58] and are associated with short dsRNAs[58]. It is possible that the steps of PTGS and TGS are, in fact, the same and differ only in their target sequences: hpRNA or dsRNA is cleaved by the plant homologue of Dicer-1 into ~21-nucleotide dsRNAs to guide specific ssRNA degradation in the cytoplasm, and a similar ribonucleoprotein complex passes into the nucleus to direct chromatin remodelling/methylation of homologous DNA (Fig. 4). Thus, production of hpRNA/dsRNA that contains promoter sequences leads to the methylation/altered state of the promoter DNA, causing TGS, whereas hpRNA/dsRNA that contains coding-region sequences leads to the degradation of homologous mRNA, causing PTGS. The methylation of coding-region DNA in PTGS and the potential degradation of promoter-sequence transcripts in TGS would be irrelevant by-products, as methylated coding regions are readily transcribed[58,59] and promoter sequences are not usually transcribed.

It seems unlikely that the DNA methylation mechanism associated with PTGS and TGS is involved directly in protecting plants against most RNA viruses. The vast majority of these viruses have exclusively cytoplasmic lifecycles and no homologous DNA sequences in plant genomes. It is possible that dsRNA-directed methylation is involved in inhibiting the handful of known plant retroviruses or pararetroviruses during their DNA phases within the nucleus. It is even more likely that the mechanism is primarily for defence against TEs.

Defence against transposons in plants

DNA methylation may have evolved as an epigenetic means of containing the spread of TEs in host genomes[4,60]. *De novo* DNA methylation was first detected in plants during the inactivation of class II TEs[61] and has been associated with both transcriptional inactivation

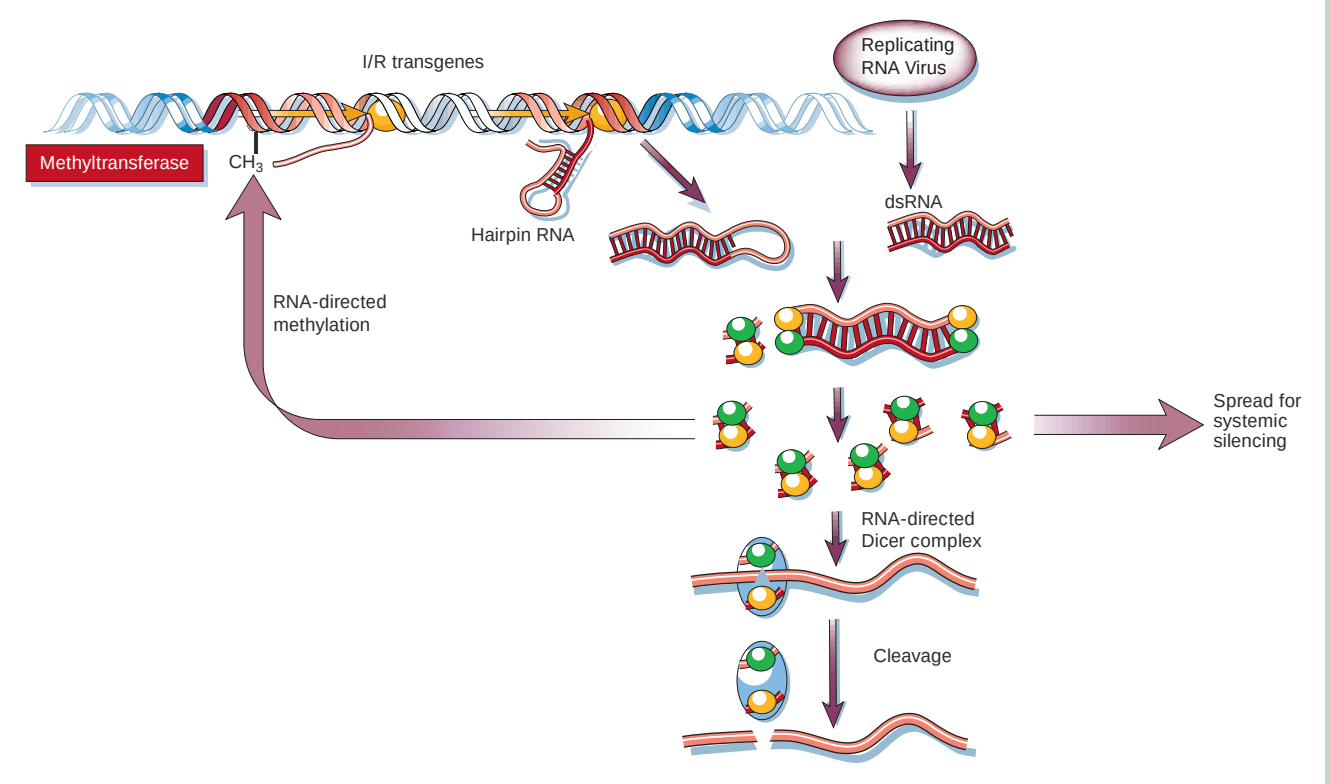


Figure 4 A model for the initiation and operation of PTGS. The hpRNA or dsRNA produced from either an inverted-repeat transgene or a replicating virus is cleaved into ~21-nucleotide fragments by the Dicer-containing complex and used as guides for cleavage of homologous ssRNA (described in more detail in Fig. 2). The ~21-nucleotide dsRNA, in some sort of complex, is transported into the nucleus to direct DNA methylation of homologous sequences or into neighbouring cells to act as a mobile PTGS signal.

© 2001 Macmillan Magazines Ltd

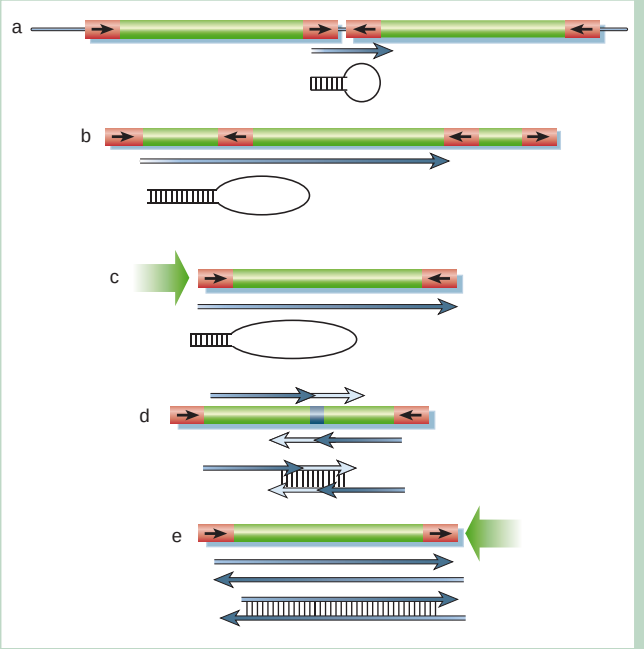


Figure 5 Possible ways in which transposons may generate hpRNA or dsRNA. **a**, LTR transposon integrated as an inverted repeat. **b**, LTR transposon integrated in the opposite polarity into another copy of the same transposon. **c**, TIR transposon adjacent to an endogenous promoter. **d**, TIR transposon (Mu). **e**, LTR transposon adjacent to an endogenous promoter. Red blocks represent either the inverted or direct terminal-repeat sequences. Green blocks represent the coding regions of the TEs. Dark blue arrows below represent the transcripts generating dsRNAs or hpRNAs, and the drawing below each transcript represents the structure it may form (that is, either a hairpin or a dsRNA structure). Blue box represents the transcription terminator sequence and light blue arrows represent readthrough transcription. Large green arrows represent endogenous promoters.

and increased transitional mutation[62–64]. Many studies have shown the involvement of methylation with transposon inactivation[61,65–67] and demethylation with transposon activation[68,69]. A recent demonstration of this comes from the study of the retrotransposon Tto1 in *Arabidopsis*[70]. As this TE becomes transcriptionally silenced, it also becomes increasingly methylated, and demethylation of the element, in a hypomethylation mutant *ddm1* background, reactivates its transcription. However, it is still not entirely clear whether the methylation itself inactivates the transposon or whether it is a secondary effect of inactivation caused by a change in chromatin structure.

Epigenetic inactivation of TEs occurs in many, possibly most, cases by its insertion into or near an already heterochromatic block of genomic DNA and the radiation of this repressed state into the elements[3]. But TEs integrating into euchromatic areas may well be the target for dsRNA-induced silencing. There are a number of scenarios (Fig. 5) of how TEs could produce dsRNA or hpRNA to trigger this mechanism. The LTRs of class I TEs contain promoter sequences, so two TE copies integrating as an inverted repeat could produce hpRNA transcripts of these sequences. TEs often integrate within each other, potentially generating transcribable, complex inverted-repeat sequences. Some transposons, such as Robertson's mutator (Mu), have convergently arranged genes which produce transcripts that, by failing to terminate or be polyadenylated at the appropriate sequences, have regions of complementarity with each other[71]. Insertion of a class II TE adjacent to an endogenous promoter directing transcription across the elements could produce RNA with self-complementarity from the TIRs. An adjacent endogenous promoter directing transcription from the reverse end of a TE could produce antisense RNA that might hybridize with the TE RNA to produce dsRNA. It is also possible that the reverse transcription of retrotransposon RNAs produces intermediates in the cytoplasm similar to replicating RNA viruses which, although RNA/DNA hybrids rather than RNA/RNA hybrids, act as triggers for dsRNA-induced silencing. Indeed, normal infection of plants with cauliflower mosaic pararetrovirus, which will produce a similar RNA/DNA intermediate, triggers a PTGS-like response[72]. But the most convincing evidence that the dsRNA-induced silencing mechanism is suppressing TEs is that such silenced elements are reactivated in a number of PTGS- and RNAi-defective mutants (Table 1), and that some of the ~21-nucleotide dsRNAs from RNAi and PTGS extracts contain sequences of TEs (ref. 40, and A. J. Hamilton and D. C. Baulcombe, personal communication). The TEs are probably controlled by methylation of their DNA and degradation of their transposase mRNA.

PTGS can spread systemically through a plant

PTGS has three phases: initiation, maintenance and, remarkably, spread[16,73]. Transgenes and viruses can initiate PTGS, as can exogenous DNA delivered by bombardment or *Agrobacterium* infiltration[16], and grafting of unsilenced scions onto silenced rootstock[73]. These last methods give localized delivery points for PTGS that spreads from these points into other tissues. It seems to spread by a non-metabolic, gene-specific diffusible signal that is capable of travelling both between cells through plasmodesmata, and long distances via the phloem[16,73]. For example, new tissue growing from a GUS-expressing scion, grafted onto a GUS-silenced rootstock, shows progressive silencing of its GUS transgene[73]. The signal seems to be sequence specific, to move uni-directionally from source to sink tissue, and can traverse at least 30 cm of wild-type stem grafted between the GUS-expressing scion and GUS-silenced rootstock[73].

To account for the specificity of the signal, it must consist (at least in part) of the transgene product, probably in the form of RNA[73]. The concept of cell-to-cell and long-distance spread of endogenous RNAs within plants remains somewhat controversial, but is not unprecedented. For instance, plant viruses have genomes composed of RNA and, when they infect their host, their RNA spreads throughout the plant. Viral-encoded movement proteins facilitate the movement of viral RNA between cells through plasmodesmata in the form of either a ribonucleoprotein complex or intact virions. To fulfil this role, movement proteins have the capacity to move between cells, bind viral RNA and dilate the size exclusion limit of plasmodesmata[74]. Simpler still, viroids — plant pathogens with small (~350 nucleotide) naked RNA genomes encoding no proteins — also infect and spread though plants, presumably associated with host proteins[75].

There is an emerging picture of RNA mobility in plants that potentially impacts on other plant processes, including transport of the gene-specific silencing signal. Examples of host RNAs moving from cell to cell include the KNOTTED1 transcription factor and its corresponding mRNA[76], and the transcript that encodes sucrose transporter 1, which has been localized to the enucleate sieve elements, presumably having been transported there from the associated companion cell[77]. Perhaps the most convincing demonstration of intercellular movement of endogenous plant RNA, and potentially signalling, is the demonstration that mRNA is found in the phloem of rice and cucurbits[78,79]. Mobility of pumpkin phloem RNA was demonstrated using grafting experiments. In one instance, a transcript encoding a transcription factor, NACP, was detected in the meristem of cucumber scions that had been heterografted onto a pumpkin rootstock[78]. Thus RNA molecules, derived from the silenced transgene, might move from cells where this gene is silenced, possibly with cellular protein factors fulfilling a role similar to viral movement proteins, to induce silencing in other cells expressing the same transgene. This raises three questions — is the signal the ~21-nucleotide dsRNA, how is the signal propagated, and what is the natural (non-transgenic) role of the signal?

© 2001 Macmillan Magazines Ltd

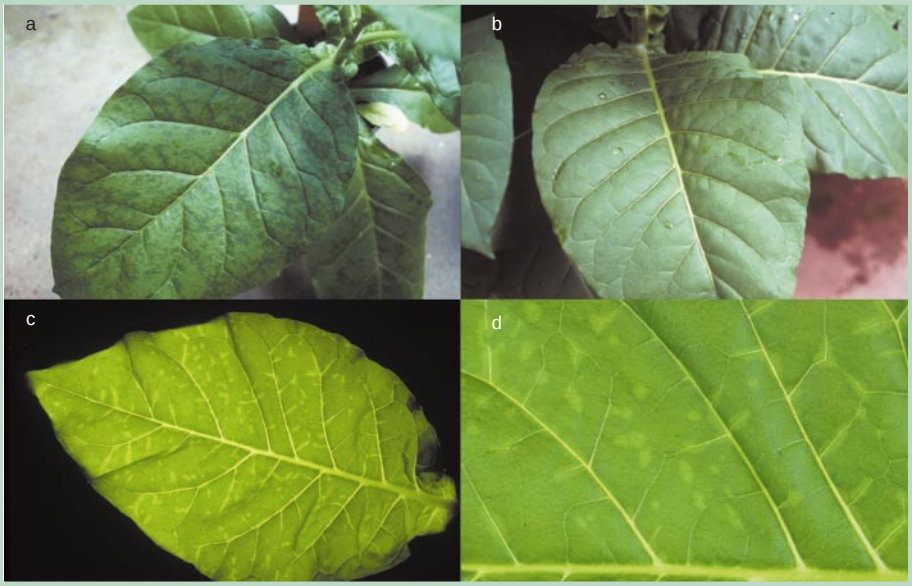


Figure 6 Tobacco plants showing potato virus Y (PVY) susceptibility, immunity and resistant/recovery symptoms. **a**, Non-transgenic tobacco plant challenged with PVY, showing a uniform chlorotic mosaic on leaves throughout the plant. **b**, Virus-challenged transgenic plant containing a transgene that expresses a hpRNA derived from PVY sequences, showing immunity to the virus and no symptoms. **c**, A whole leaf, and **d**, close-up, from a PVY-challenged transgenic plant showing resistance/recovery symptoms. As the plant grows the leaves show fewer and fewer yellow patches; only the yellow patches contain detectable levels of virus. The pattern suggests that as the virus attempts to spread from the phloem into the surrounding cells, a signal is emitted that allows more distal cells to be forewarned and resist and restrict the virus.

Defence is a two-step process

Most plant species are immune to most plant viruses, possibly due to compatibility requirements between the RDRP or movement protein of a specific virus and the host factors present in the plant. However, it is also possible that the only virus/host combinations leading to an infection are ones in which the virus can overcome or avoid the plant's PTGS defence mechanism. For example, potyviruses encode a protein, HC-Pro, that potently disables PTGS[31,80,81] by directly or indirectly preventing the dsRNA cleaving activity of the Dicer complex. Therefore, a plant challenged with a potyvirus becomes a battleground in the fight between a defence and a counter-defence strategy. It is a race between the cell's recognition and degradation of viral RNA, and the virus expressing HC-Pro to inactivate the degradation machinery. Perhaps the mobile signal, seen in the artificial situation of transgenes and grafting experiments, is a reflection of a plant's fallback strategy. If the first-challenged cells are not quick enough to recognize and destroy the virus, but can send a warning message to uninfected cells of their imminent invasion, these cells can have their degradation machinery prepared. If the signal contains fragments of the virus sequence, the recipient cells can be ready to degrade RNAs containing these sequences before the virus arrives (Fig. 6). This model might explain a number of observations: (1) plants transformed with virus-derived transgenes can show a recovery response (that is, initial virus infection and symptoms followed by new growth that is both without symptoms and resistant to virus infection); (2) infection of non-transgenic plants with nepo-, tobra- and caulimoviruses can induce responses very much like the recovery phenotype[72,82,83]; and (3) plants showing co-suppression tend to initially show activity of the target gene, but this becomes progressively more silenced in developing tissue. All of these progressive silencing phenotypes could be the result of cells that induce PTGS too slowly or poorly to significantly degrade the virus or transgene RNA, but produce a signal that is amplified in recipient cells such that they can perform effective silencing.

Viruses are capable of rapid evolution to overcome the plant's host defences, even their fallback plans, wherever they can. So it is of little surprise that at least two genera of plant viruses (potex- and cucumoviruses) overcome their host's PTGS mechanism by producing proteins that seem to prevent the spread of the signal[84–86].

What is the signal?

Initial impressions might suggest that the ~21-nucleotide dsRNAs produced by Dicer not only direct degradation, but are also the signal (Fig. 4) that moves into the nucleus to direct DNA methylation and that moves from cell to cell to be amplified and direct RNA degradation and DNA methylation there. But there is strong evidence from grafting experiments, which exploit HC-Pro, that these short dsRNAs are not the signal[87]. As discussed above, the expression of HC-Pro restores activity to transgenes previously showing PTGS and prevents the production of short dsRNAs. In addition, the transgene in a GUS-expressing scion grafted onto a GUS-silenced rootstock becomes silenced. However, when a GUS-expressing scion is grafted onto a rootstock in which a once-silenced GUS transgene has been reactivated by HC-Pro, GUS activity in the scion is silenced. In this grafted plant, short RNAs are detected in the scion, but not in the rootstock. This indicates that short RNAs are not the mobile signal, as they were not detected in the rootstock of the graft, yet the rootstock transmitted a silencing signal to the scion. In addition, once-silenced GUS transgenes, which have been reactivated by HC-Pro (with the concomitant absence of short RNAs), still develop PTGS-like methylation. This casts doubt on the idea that short RNAs act as guides for methylation, although other workers have found that inhibition of silencing and short RNA production by HC-Pro reduced the level of target gene methylation[31].

HC-Pro probably interacts in the PTGS pathway upstream of the production of short RNAs[31,87], but downstream of the production of the signal. So what might the signal molecule be, if not ~21-nucleotide dsRNA with or without Dicer-1? It might be a modified product from the methylated target gene[87] or dsRNA itself. Perhaps dsRNA is recognized in a cell by both the Dicer complex and by another complex containing RNA-movement proteins (Fig. 7). This latter complex facilitates the passage of the dsRNA through nuclear pores and plasmodesmata. In the nucleus, the dsRNA is unwound and used to direct methylation. The dsRNA entering cells, in which the PTGS mechanism is not yet activated (such as in a graft situation), attracts degradation complexes which then produce short RNAs. These act as primers on target mRNAs for a host RDRP, for example, SGS2 (perhaps in association with AGO1, and the helicase SDE3), to produce complementary strands of RNA, thus making dsRNA. The dsRNA is then the substrate for both the movement and PTGS-degradation complexes, thereby re-amplifying the mobile signal and propagating PTGS. This scenario would make the plant's RNA-movement protein and RDRP (and associated factors, including an RNA helicase) essential for amplifying weak signals into effective silencing. Evidence supporting this primer model is that the signal seems to amplify from the target mRNA. GFP-expressing plants silenced by bombardment with DNA encoding one-third of the GFP sequence are resistant to a virus containing the other two-thirds of

© 2001 Macmillan Magazines Ltd

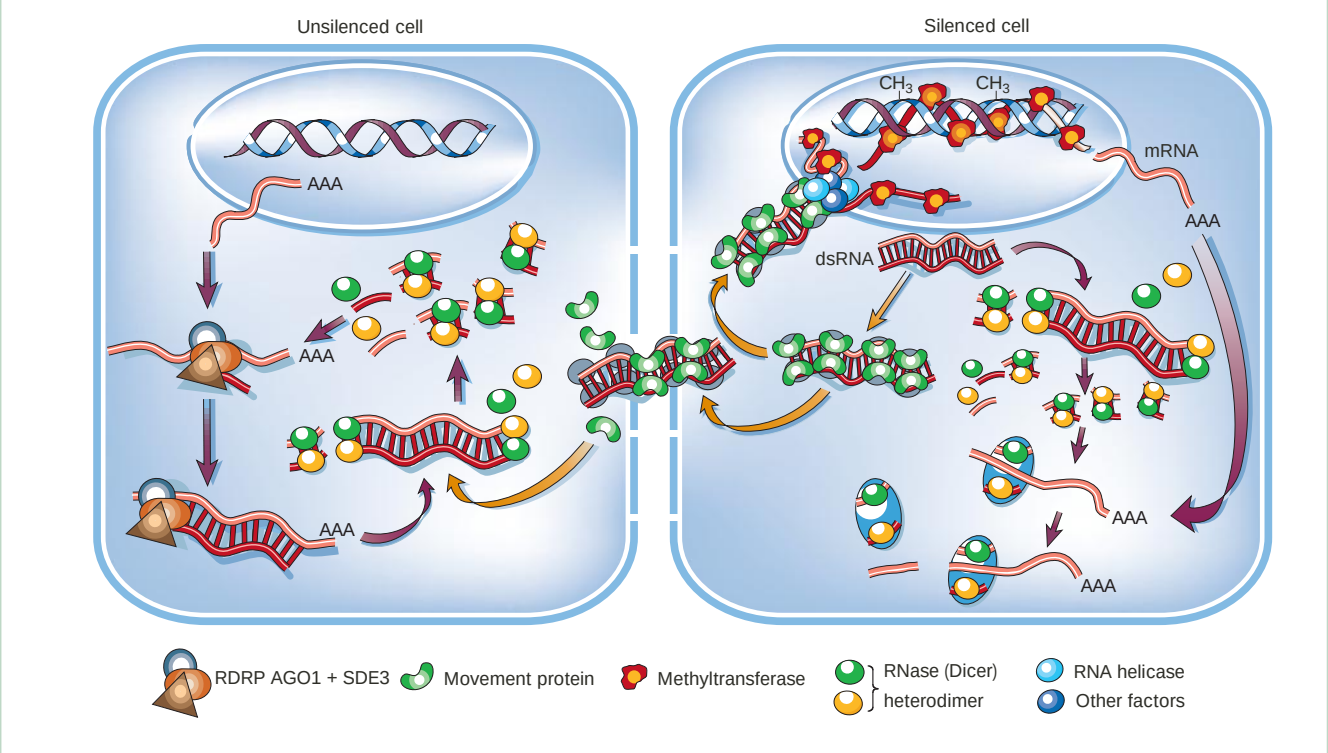


Figure 7 A model for mobile PTGS and methylation signals. The dsRNA-induced cleavage of ssRNA is described in Figs 2 and 4. However, an extra dsRNA-specific movement-protein complex may be required for the spread and amplification of the PTGS signal. This complex transports the dsRNA into the nucleus via the nuclear pore, and into adjacent cells (and ultimately the phloem) by means of the plasmodesmata. In the nucleus, the dsRNA is unwound by a helicase and directs methylation of homologous DNA sequences. In situations such as the grafting of an unsilenced scion onto a silenced rootstock, the dsRNA complex entering the unsilenced cell of the scion is recognized and degraded, releasing some dissociated short RNAs. These RNAs act as primers on a target mRNA for a host RDRP (for example, SGS2) with a helicase (for example, SDE3) and initiating factors (for example, AGO1) to synthesize complementary RNA. This generates dsRNA that is recognized by the components of the movement and degradation complexes, thus initiating degradation of homologous ssRNA and amplifying the mobile signal.

the GFP gene sequence[16]. Because there is no overlap between these GFP-derived sequences, it seems that the PTGS targeted against the GFP-containing virus was mediated through the mRNA (or DNA) of the intact GFP transgene in the plant.

The picture so far

As some of the parts of this jigsaw puzzle come together, a picture begins to emerge. A number of the pieces seem to fit together perfectly, some not so perfectly, and many pieces are still missing. The pieces describing the mechanism of dsRNA induction and directed ssRNA degradation fit together well. Those describing the directed methylation and signal amplification fit into the picture but have quite a few adjacent pieces missing. SGS3 and RDE4 are like pieces whose colour and pattern tell us they are part of the puzzle, but as yet they remain unconnected. However, the overall picture is of an ingenious defence mechanism in plants against viruses and TEs. That PTGS is a manifestation of an anti-virus defence mechanism seems beyond doubt from the discovery that viruses encode proteins to specifically disable the system[81,84–86]. There is considerable evidence that RNAi in nematodes is a defence mechanism against TEs. Out of 30 mutants that activate transposons in nematodes, 22 also cause defects in RNAi[49]. It seems highly likely that homologous genes have the same role in plants.

This plant defence mechanism against viruses and TEs has a number of parallels with the immune system of mammals. The immune system detects foreign bodies and generates agents (antibodies) to specifically recognize them and guide destruction machinery (phagocytes and killer cells) against them. The plant's defence system generates agents (short dsRNAs) with the specificity to recognize sequences in the invading viral RNA genomes and guide destruction complexes (nucleases) to them. Both systems work systemically, viruses have responded to both systems by evolving ways to circumvent them, and both systems can be enhanced by prior vaccination with virus components.

This plant defence system is an elegant example of how nature can find highly efficient solutions to the problems it faces. Just as translation machinery needs codons of only 3 nucleotides to specify which amino acid goes where in a protein, this defence system requires degradation recognition sequences of only ~21 nucleotides to direct it. With each 21-nucleotide guide sequence, the probability of targeting an inappropriate 1-kilobase mRNA is about 1 in 10^{10}. The minimum length of dsRNA needed to induce the system is probably ~21 nucleotides, which must place evolutionary constraints on the self-complementarity of mRNAs. Indeed, codon redundancy may be exploited in mRNAs to avoid perfect self-complementarity regions of greater than 20 nucleotides. In addition, plants might use self-complementary motifs to give added instability features to a mRNA. It is possible that chalcone synthase mRNA in plants, which has regions of significant self-complementarity, might exploit this mechanism to generate stochastic pigmented patterns in petals to attract insects (and plant breeders). The PTGS mechanism may be involved in controlling plant development. AGO1, and possibly CAF1, is needed for PTGS; mutants in either of these genes cause developmental abnormalities[48,88]. It is possible that PTGS controls development by regulating the abundance or transport of endogenous RNAs in plants. Interestingly, a plant host protein, rgs-CaM, has been found which, when overexpressed, disables PTGS in a similar manner to HC-Pro[89]. Perhaps the cell uses this protein to temporarily inactivate PTGS at appropriate times in the cell cycle or in developmental states when PTGS could be detrimental.

© 2001 Macmillan Magazines Ltd

We are beginning to understand that phenotypes such as co-suppression and virus-derived transgene-mediated resistance are manifestations of a plant defence mechanism. We also know some design rules[20] of how to exploit the mechanism to generate plants with traits for commercial agriculture (for example, virus protection[90] and modification of quality traits by inhibiting steps of metabolic pathways) and for gene discovery[26] (for example, specifically silencing genes of unknown function in the *Arabidopsis* genome database). But many aspects remain to be elucidated and there are some seemingly contradictory results to be reconciled. It is also intriguing that although directed RNA degradation and DNA methylation are aspects of the plant's defence mechanism, these processes may also be important in whole-plant development and homeostasis mediated by RNA-based intercellular communication. □

1. Mayo, M. A. Developments in plant virus taxonomy since the publication of the 6th ICTV report. *Arch. Virol.* **144,** 1659–1666 (1999).
2. Bennetzen, J. L. Transposable element contributions to plant gene and genome evolution. *Plant Mol. Biol.* **42,** 251–269 (2000).
3. Kumar, A. & Bennetzen, J. L. Plant retrotransposons. *Annu. Rev. Genet.* **33,** 479–452 (1999).
4. Martienssen, R. Transposons, DNA methylation and gene control. *Trends Genet.* **14,** 263–264 (1998).
5. Rothstein, S. J., Dimaio, J., Strand, M. & Rice, D. Stable and inheritable inhibition of the expression of nopaline synthase in tobacco expressing antisense RNA. *Proc. Natl Acad. Sci. USA* **84,** 8439–8443 (1987).
6. Simons, R. W. & Kleckner, N. Translational control of IS10 transposition. *Cell* **34,** 683–691 (1983).
7. Hamilton, A. J., Lycett, G. W. & Grierson, D. Antisense gene that inhibits synthesis of the hormone ethylene in transgenic plants. *Nature* **346,** 284–287 (1990).
8. Napoli, C., Lemieux, C. & Jorgensen, R. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. *Plant Cell* **7,** 599–609 (1990).
9. Van der Krol, A. R., Mur, L. A., Beld, K., Mol, J. N. M. & Stuitje, A. R. Flavonoid genes in petunia: addition of a limited number of gene copies may lead to suppression of gene expression. *Plant Cell* **2,** 291–299 (1990).
10. De Carvalho, F. *et al.* Suppression of the β-1,3 glucanase transgene expression in homozygous plants. *EMBO J.* **11,** 2595–2602 (1992).
11. Powell-Abel, P. *et al.* Delay of disease development in transgenic plants that express the tobacco mosaic virus coat protein gene. *Science* **232,** 738–743 (1986).
12. Lindbo, J. A. & Dougherty, W. G. Untranslatable transcripts of the tobacco etch virus coat protein gene sequence can interfere with tobacco etch virus replication in transgenic plants and protoplasts. *Virology* **189,** 725–733 (1992).
13. Dougherty, W. G. & Parks, T. D. Transgenes and gene suppression: telling us something new? *Curr. Opin. Cell Biol.* **7,** 399–405 (1995).
14. English, J. J., Mueller, E. & Baulcombe, D. C. Suppression of virus accumulation in transgenic plants exhibiting silencing of nuclear genes. *Plant Cell* **8,** 787–797 (1996).
15. Tanzer, M. M., Thompson, W. F., Law, M. D., Wernsman, E. A. & Ukenes, S. Characterization of post-transcriptionally suppressed transgene expression that confers resistance to tobacco etch virus infection in tobacco. *Plant Cell* **9,** 1411–1423 (1997).
16. Voinnet, O., Vain, P., Angell, S. & Baulcombe, D. C. Systemic spread of sequence-specific transgene RNA degradation in plants is initiated by localized introduction of ectopic promoterless DNA. *Cell* **95,** 177–187 (1998).
17. Wassenegger, M. & Pelissier, T. A model for RNA-mediated gene silencing in higher plants. *Plant Mol. Biol.* **37,** 349–362 (1998).
18. Waterhouse, P. M., Graham, M. W. & Wang, M.-B. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc. Natl Acad. Sci USA* **95,** 13959–13964 (1998).
19. Wang, M.-B. & Waterhouse, P. M. High efficiency silencing of a β-glucuronidase gene in rice is correlated with repetitive transgene structure but is independent of DNA methylation. *Plant Mol. Biol.* **43,** 67–82 (2000).
20. Smith, N. A. *et al.* Total silencing by intron-spliced hairpin RNAs. *Nature* **407,** 319–320 (2000).
21. Stam, M. *et al.* Distinct features of post-transcriptional gene silencing by antisense transgenes in single copy and inverted T-DNA repeat loci. *Plant J.* **21,** 27–42 (2000).
22. Morris T. J. & Dodds, J. A. Isolation and analysis of double stranded RNA from virus-infected plant and fungal tissue. *Phytopathology* **69,** 854–858 (1979).
23. Blumenthal, T. & Carmichael, G. G. RNA replication: function and structure of Qβ replicase. *Annu. Rev. Biochem.* **48,** 525–548 (1979).
24. Buck, K. W. Replication of tobacco mosaic virus RNA. *Phil. Trans. R. Soc. Lond. B* **354,** 613–627 (1999).
25. Ogawa, T., Hori, T. & Ishida, I. Virus-induced death in plants expressing the mammalian 2′,5′ oligoadenylate system. *Nature Biotechnol.* **14,** 1566–1569 (1996).
26. Baulcombe, D. C. Fast forward genetics based in virus-induced gene silencing. *Curr. Opin. Plant Biol.* **2,** 109–113 (1999).
27. Lee, K. Y., Baden, C., Howie, W. J., Bedbrook, J. & Dunsmuir, P. Post-transcriptional gene silencing of ACC synthase in tomato results from cytoplasmic RNA degradation. *Plant J.* **12,** 1127–1137 (1997).
28. Metzlaff, M., O'Dell, M., Cluster, P. D. & Flavell, R. B. RNA-mediated RNA degradation and chalcone synthase A silencing in *Petunia*. *Cell* **88,** 845–854 (1997).
29. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286,** 950–952 (1999).
30. Mette, M. F., Aufatz, W., van der Winden, J., Matzke, M. A & Matzke, A. J. M. Transcriptional silencing and promoter methylation triggered by double stranded RNA. *EMBO J.* **19,** 5194–5201 (2000).
31. Llave, C., Kasschau, K. D. & Carrington, J. C. Virus-encoded suppressor of posttranscriptional gene silencing targets a maintenance step in the silencing pathway. *Proc. Natl Acad. Sci. USA* **97,** 13401–13406 (2000).
32. Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in Arabidopsis is required for post-transcriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101,** 543–553 (2000).
33. Hutvagner, G., Mlynarova, L. & Nap, J. P. Detailed characterization of the posttranscriptional gene-silencing-related small RNA in a GUS gene-silenced tobacco. *RNA* **6,** 1445–1454 (2000).
34. Fire, A. *et al.* Potent and specific genetic interference by double stranded RNA in *Caenorhabditis elegans*. *Nature* **391,** 806–811 (1998).
35. Kennerdel, J. R. & Carthew, R. W. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell* **95,** 1017–1026 (1998).
36. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101,** 25–33 (2000).
37. Yang, D., Lu, H. & Erickson, J. W. Evidence that processed small dsRNAs may mediate sequence-specific mRNA degradation during RNAi in *Drosophila* embryos. *Curr. Biol.* **10,** 1191–1200 (2000).
38. Hammond, S. C., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404,** 293–296 (2000).
39. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409,** 363–366 (2001).
40. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.* **15,** 188–200 (2001).
41. Holtorf, H., Schob, H., Kunz, C., Waldvogel, R. & Meins, F. Stochastic and nonstochastic post-transcriptional silencing of chitinase and β-1,3-glucanase genes involves increased RNA turnover—possible role for ribosome-independent RNA degradation. *Plant Cell* **11,** 471–483 (1999).
42. Zou, C., Zhang, Z., Wu, S. & Osterman, J. C. Molecular cloning and characterization of a rabbit eIF2C protein. *Gene* **211,** 187–194 (1998).
43. Cogoni, C. & Macino, G. Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase. *Nature* **399,** 166–169 (1999).
44. Mourrain, P. *et al. Arabidopsis SGS2* and *SGS3* genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101,** 533–542 (2000).
45. Smardon, A. *et al.* EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in *C. elegans*. *Curr. Biol.* **10,** 169–178 (2000).
46. Catalanotto, C., Azzalin, G., Macino, G. & Cogoni, C. Gene silencing in worms and fungi. *Nature* **404,** 245 (2000).
47. Tabara, H. *et al.* The rde-I gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99,** 123–132 (1999).
48. Fagard, M., Boutet, S., Morel, J.-B., Bellini, C. & Vaucheret, H. AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi and RNA interference in animals. *Proc. Natl Acad. Sci. USA* **97,** 11650–11654 (2000).
49. Ketting, R. F., Haverkamp, T. H. A., van Luenen, H. G. A. & Plasterk, R. H. A. *mut-7* of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner Syndrome helicase and RNase D. *Cell* **99,** 133–141 (1999).
50. Domeier, M. E. *et al.* A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*. *Science* **289,** 1928–1930 (2000).
51. Cogoni, C. & Macino, G. Post-transcriptional gene silencing in *Neurospora* by a ReqQ DNA helicase. *Science* **286,** 342–344 (1999).
52. Wu-Scarf, D., Jeong, B.-R., Zhang, C. & Cerutti, H. Transgene and transposon silencing in *Chlamydomonas reinhardtii* by a DEAH-box RNA helicase. *Science* **290,** 1159–1162 (2000).
53. Dalmay, T., Horsefield, R., Braunstein, T. H. & Baulcombe, D. C. SDE3 encodes an RNA helicase required for post-transcriptional gene silencing in *Arabidopsis*. *EMBO J.* **20,** 2069–2077 (2001).
54. Cerruti, L., Mian, N. & Bateman, A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25,** 481–482 (2000).
55. Kooter, J. M., Matzke, M. A. & Meyer, P. Listening to the silent genes: transgene silencing, gene regulation and pathogen control. *Trends. Plant Sci.* **4,** 340–347 (1999).
56. Morel, J.-B., Mourrain, P., Berlin, C. & Vaucheret, H. DNA methylation and chromatin structure mutants affect both post-transcriptional and transcriptional transgene silencing in *Arabidopsis*. *Curr. Biol.* **10,** 1591–1594 (2001).
57. Jones, A. L. *et al. De novo* methylation and co-suppression induced by a cytoplasmically replicating plant RNA virus. *EMBO J.* **17,** 6385–6393 (1998).
58. Wang, M.-B., Wesley, S. V., Finnegan, E. J., Smith, N. A. & Waterhouse, P. M. Replicating satellite RNA induces sequence specific DNA methylation and truncated transcripts in plants. *RNA* **7,** 16–28 (2001).
59. Van Houdt, H., Ingelbrecht, I., van Montagu, M. & Depicker, A. Post-transcriptional silencing of a neomycin phosphotransferase II transgene correlates with the accumulation of unproductive RNAs and with increased cytosine methylation of 3′ flanking regions. *Plant J.* **12,** 379–392 (1997).
60. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and ecology of intragenomic parasites. *Trends Genet.* **13,** 335–340 (1997).
61. Brutnell, T. P. & Dellaporta, S. L. Somatic inactivation and reactivation of Ac associated changes in cytosine methylation and transposase expression. *Genetics* **138,** 213–225 (1994).
62. San Miguel, P. *et al.* Nested transposons in the intergenic regions of the maize genome. *Science* **257,** 765–768 (1996).
63. Bennetzen, J. L. The contributions of retroelements to plant genome organisation, function and evolution. *Trends Microbiol.* **4,** 347–353 (1993).
64. Kunze, R., Saedler, H. & Lonnig, W. E. Plant transposable elements. *Adv. Bot. Res.* **27,** 331–470 (1997).
65. Bennetzen, J. L. The *Mutator* transposable element system of maize. *Curr. Top. Microbiol. Immunol.* **204,** 195–229 (1996).
66. Schlappi, M., Raina, R. & Fedoroff, N. Epigenetic regulation of the maize Spm transposable element: novel activation of a methylated promoter by TnpA. *Cell* **77,** 427–437 (1994).
67. Liu, W. M., Maraia, R. J., Rubin, C. M. & Schmid, C. W. Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. *Nucleic Acids Res.* **22,** 1087–1095 (1994).
68. Singer, T., Yordan, C. & Martienssen, R. A. Robertson's Mutator transposons in *A. thaliana* are regulated by the chromatin-remodeling gene Decrease in DNA Methylation (DDM1). *Genes Dev.* **15,** 591–602 (2001).
69. Miura, A. *et al.* Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature* **411,** 212–214 (2001).
70. Hirochika, H., Okamoto, H. & Kakutani, T. Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. *Plant Cell* **12,** 357–368 (2000).
71. Hershberger, R. J. Characterization of the major transcripts encoded by the regulatory MuDR transposable element of maize. *Genetics* **140,** 1087–1098 (1995).

© 2001 Macmillan Magazines Ltd

72. Covey, S. N., Al-Kaff, N. S., Langara, A. & Turner, D. S. Plants combat infection by gene silencing. *Nature* **385,** 781–782 (1997).
73. Palauqui, J. C., Elmayan, T., Pollien, J. M. & Vaucheret, H. Systemic acquired silencing: transgene specific post transcriptional silencing is transmitted by grafting from silenced stocks to non-silenced scions. *EMBO J.* **16,** 4738–4745 (1997).
74. Lucas, W. J. & Gilbertson, R. L. Plasmodesmata in relation to viral movement within leaf tissues. *Annu. Rev. Phytopathol.* **32,** 387–411 (1994).
75. Ding, B., Kwon, M. O., Hammond, R. & Owens, R. Cell to cell movement of potato spindle tuber viroid. *Plant J.* **12,** 931–936 (1997).
76. Lucas, W. J. *et al.* Selective trafficking of KNOTTED1 homoeodomain protein and its mRNA through plasmodesmata. *Science* **270,** 1980–1983 (1995).
77. Kuhn, C., Franceschi, V. R., Schulz, A., Lemoine, R. & Frommer, W. B. Macromolecular trafficking indicated by localisation and turnover of sucrose in enucleate sieve elements. *Science* **275,** 1298–1300 (1997).
78. Ruiz-Medrano, R., Xocosnostle-Cazares, B. & Lucas, W. J. Phloem long-distance transport of CmNACP mRNA: implications for supracellular regulation in plants. *Development* **126,** 4405–4419 (1999).
79. Sasaki, T., Chino, M., Hayashi, H. & Fujiwara, T. Detection of mRNA species in phloem sap. *Plant Cell Physiol.* **39,** 895–897 (1998)
80. Kasschau, K. D. & Carrington J. C. A counter-defensive strategy of plant viruses: suppression of post-transcriptional gene silencing. *Cell* **95,** 461–470 (1998).
81. Marathe, M. *et al.* Plant viral suppressors of post-transcriptional silencing do not suppress transcriptional silencing. *Plant J.* **22,** 51–59 (2000).
82. Ratcliff, F., Harrison, B. D. & Baulcombe, D. C. A similarity between viral defense and gene silencing in plants. *Science* **276,** 1558–1560 (1997).
83. Ratcliff, F. G., MacFarlane, S. S. & Baulcombe, D. C. Gene silencing without DNA: RNA mediated cross protection between viruses. *Plant Cell* **11,** 1207–1215 (1999).
84. Brigneti, G. *et al.* Viral pathogenicity determinants are suppressors of transgene silencing in *Nicotiana bentamiana*. *EMBO J.* **17,** 6739–6746 (1998).
85. Li, H. W. *et al.* Strong host resistance targeted against a viral suppressor of the plant gene silencing defence mechanism. *EMBO J.* **18,** 2683–2691 (1999).
86. Voinnet, O., Lederer, C. & Baulcombe, D. C. A viral movement protein prevents spread of the gene silencing signal in *Nicotiana benthamiana*. *Cell* **103,** 157–167 (2000).
87. Mallory, A. C. *et al.* HC-Pro suppression of gene silencing eliminates the small RNAs but not the transgene methylation or the mobile signal. *Plant Cell* **13,** 571–583 (2001).
88. Jacobsen, S. E., Running, M. P. & Meyerowitz, E. M. Disruption of an RNA helicase/RNaseIII gene in Arabidopsis causes unregulated cell division in floral meristems. *Development* **126,** 5231–5243 (1999).
89. Anandakshmi, R. *et al.* A calmodulin-related protein that suppresses posttranscriptional gene silencing in plants. *Science* **290,** 142–144 (2000).
90. Wang, M.-B., Abbott, D. & Waterhouse, P. M. A single copy of a virus-derived transgene encoding hairpin RNA gives immunity to barley yellow dwarf virus. *Mol. Plant Pathol.* **1,** 347–356 (2000).

Acknowledgements
We thank M.-A. Grandbastion, R. Hull and our colleagues at P.I. for being so generous with their time and knowledge. Thanks also to V. Vance for access to her manuscript before publication.

© 2001 Macmillan Magazines Ltd

Natural products and plant disease resistance

Richard A. Dixon

Plant Biology Division, Samuel Roberts Noble Foundation, Ardmore, Oklahoma 73401, USA

Plants elaborate a vast array of natural products, many of which have evolved to confer selective advantage against microbial attack. Recent advances in molecular technology, aided by the enormous power of large-scale genomics initiatives, are leading to a more complete understanding of the enzymatic machinery that underlies the often complex pathways of plant natural product biosynthesis. Meanwhile, genetic and reverse genetic approaches are providing evidence for the importance of natural products in host defence. Metabolic engineering of natural product pathways is now a feasible strategy for enhancement of plant disease resistance.

Collectively, plants produce a remarkably diverse array of over 100,000 low-molecular-mass natural products, also known as secondary metabolites. Secondary metabolites are distinct from the components of intermediary (primary) metabolism in that they are generally non-essential for the basic metabolic processes of the plant. Most are derived from the isoprenoid, phenylpropanoid, alkaloid or fatty acid/polyketide pathways (Fig. 1). This rich diversity results in part from an evolutionary process driven by selection for acquisition of improved defence against microbial attack or insect/animal predation. However, such diversity has made it difficult to apply conventional molecular and genetic techniques to address the functions of natural products in plant defence, or to improve plant disease resistance using metabolic pathway engineering.

Related plant families generally make use of related chemical structures for defence (for example, isoflavonoids in the Leguminosae, sesquiterpenes in the Solanaceae), although some chemical classes are used for defensive functions across taxa (for example, phenylpropanoid derivatives). Some species produce a broad range of antimicrobial compounds. For example, cocoa, when infected by the vascular wilt fungus *Verticillium dahliae*, accumulates the pentacyclic triterpene arjunolic acid, two hydroxylated acetophenones and, most unusually, elemental sulphur (Fig. 1, **25**), the only known inorganic antimicrobial agent produced by plants[1]. Most antimicrobial plant natural products have relatively broad-spectrum activity, and specificity is often determined by

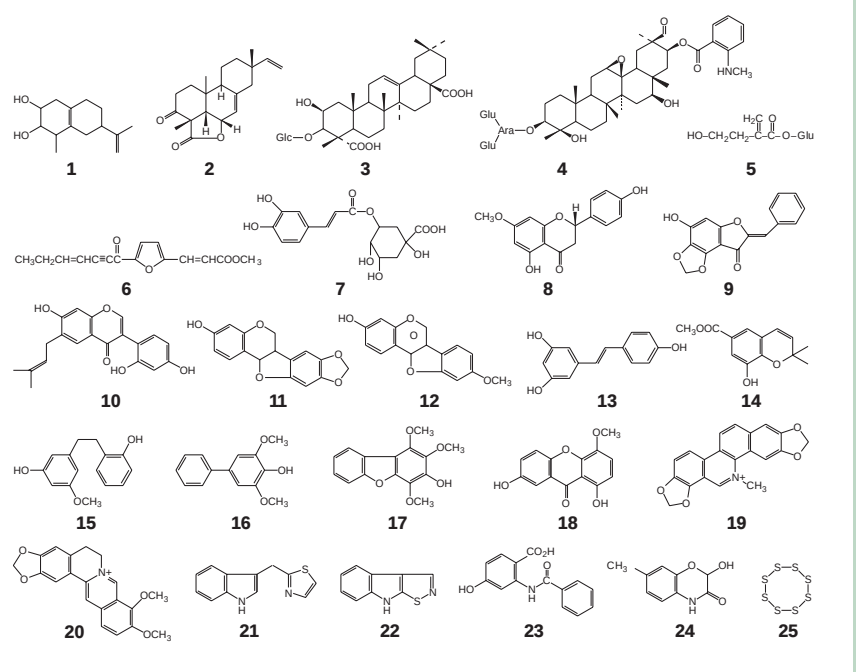


Figure 1 Chemical diversity of constitutive and inducible antimicrobial plant natural products. The chemical class of the compound is followed in parentheses by the trivial name (if available), a selected species of origin, and an indication as to whether the compound is produced constitutively (C) or is inducible (I). Terpenoids: **1**, sesquiterpene (rishitin, *Nicotiana tabacum*, I); **2**, diterpene (momilactone A, *Oryza sativa*, I); **3**, saponin (medicagenic acid 3-*O*-glucoside, *Dolichos kilimandscharicus*, C); **4**, saponin (avenacin A, *Avena sativa*, C). Aliphatic acid derivatives: **5**, butyrolactone precursor (tuliposide A, *Tulipa* spp., C); **6**, furanoacetylene (wyerone, *Vicia faba*, I). Phenolics and phenylpropanoids: **7**, hydroxycinnamic acid ester (chlorogenic acid, *Nicotiana tabacum*, C); **8**, flavanone (sakuranetin, *Ribes nigrum*, C; *Oryza sativa*, I); **9**, aurone (*Cephalocereus senilis*, I); **10**, isoflavone (luteone, *Lupinus albus*, C); **11**, pterocarpan (maackiain, *Cicer arietinum*, I); **12**, pterocarpan (medicarpin, *Medicago sativa*, I); **13**, stilbene (resveratrol, *Vitis viniferis*, I); **14**, chromene (*Piper aduncum*, C); **15**, bibenzyl (batatasin IV, *Dioscorea batatas*, C), **16**, biphenyl (aucuparin, *Malus pumila*, I); **17**, benzofuran (*Cotoneaster* spp., I); **18**, xanthone (*Polygala nyikensis*, C). Nitrogen- and/or sulphur-containing compounds: **19**, benzophenanthridine alkaloid (sanguinarine, *Papaver bracteatum*, I); **20**, benzylisoquinoline alkaloid (berberine, *Berberis* spp., I); **21**, indole (camalexin, *Arabidopsis thaliana*, I); **22**, indole (brassilexin, *Brassica* spp., I); **23**, anthranilamide (*Dianthus caryophyllus*, I); **24**, benzoxazinone (DIMBOA, *Zea mays*, C); **25**, elemental sulphur (*Theobroma cacao*, I).

© 2001 Macmillan Magazines Ltd

whether or not a pathogen has the enzymatic machinery to detoxify a particular host product[2]. Accumulation of inducible antimicrobial compounds is often orchestrated through signal-transduction pathways linked to perception of the pathogen by receptors encoded by host resistance genes (see review by Dangl and Jones, pages 826–833).

Arabidopsis, rice, corn, soya bean and the model legume *Medicago truncatula*, which have been subject to intensive sequencing efforts, are, collectively, rich sources of antimicrobial indole, terpenoid, benzoxazinone and flavonoid/isoflavonoid natural products. Bioinformatic analysis of large-scale plant genomic and expressed sequence tag (EST) databases[3,4] is beginning to reveal how new enzymes of natural product biosynthesis may have arisen through processes of gene duplication and mutation[5]. This provides the genetic variation that leads to continued elaboration of new chemical structures that will be selected for if they impart a significant advantage in plant defence.

Phytoalexins and phytoanticipins

Details of the structures and sources of many antimicrobial plant natural products have been compiled[6,7], and evidence for the functions of these compounds has also been reviewed[8,9]. The simplest functional definitions recognize phytoalexins as compounds that are synthesized *de novo* (as opposed to being released by, for example, hydrolytic activity) and phytoanticipins as pre-formed infectional inhibitors[10]. However, the distinction between phytoalexin and phytoanticipin is not always obvious. Some compounds may be phytoalexins in one species and phytoanticipins in others. A good example is the methylated flavanone sakuranetin (Fig. 1, **8**), which accumulates constitutively in leaf glands of blackcurrant, but which is a major inducible antimicrobial metabolite in rice leaves[11]. In cases where a constitutive metabolite is produced in larger amounts after infection, its status as a phytoalexin would depend on whether or not the constitutive concentrations were sufficient to be antimicrobial.

In vivo antimicrobial activity is inherent in the definition of a phytoalexin or phytoanticipin, but it is this feature that has proven most problematical to determine directly in the absence of methods to genetically modify the host plant's natural product profiles. In most cases, concentrations of phytoalexins have not been measured specifically in the cells that are in direct contact with the invading microorganism. One exception is a careful study of the cellular concentrations of sesquiterpenoid phytoalexins in leaves of cotton varieties responding to the bacterial pathogen *Xanthomonas campestris* pv. *malvacearum*, in which it was shown that phytoalexin levels in and around the challenged cells were significantly higher than would be required to effectively inhibit the growth of the pathogen *in vitro*[12].

Genetic approaches to natural product function

One approach for addressing phytoalexin or phytoanticipin function *in vivo* has been to take advantage of the power of gene-knockout technology in microbes to disrupt pathogen genes that encode enzymes known to detoxify the host plant's antimicrobial compounds. For example, saponins are widely occurring, constitutively expressed, glycosylated steroids, steroidal alkaloids or triterpenes, many of which have antimicrobial activity *in vitro*. Mutants of the fungal pathogen of oat roots, *Gaeumannomyces graminis*, that had lost the enzyme avenacinase were no longer able to detoxify the triterpene saponin avenacin (Fig. 1, **4**); they lost pathogenicity on oats, but retained full pathogenicity on wheat, which does not produce saponins[13]. This indicates that avenacin, which is localized to the epidermal cells that would represent one of the first barriers to infection by *G. graminis*, is a key determinant of resistance in this particular host–pathogen system. Likewise, disruption of the *MAK1* gene in the fungal pathogen *Nectria haematococca* leads to inability to detoxify the pterocarpan phytoalexin maackiain (Fig. 1, **11**) and reduced virulence of the fungus on chickpea. In this case the effect was incomplete, and host factors in addition to maackiain are therefore involved in resistance[14].

In a few cases, a role for plant natural products in host defence has been demonstrated by the increased disease susceptibility of mutants

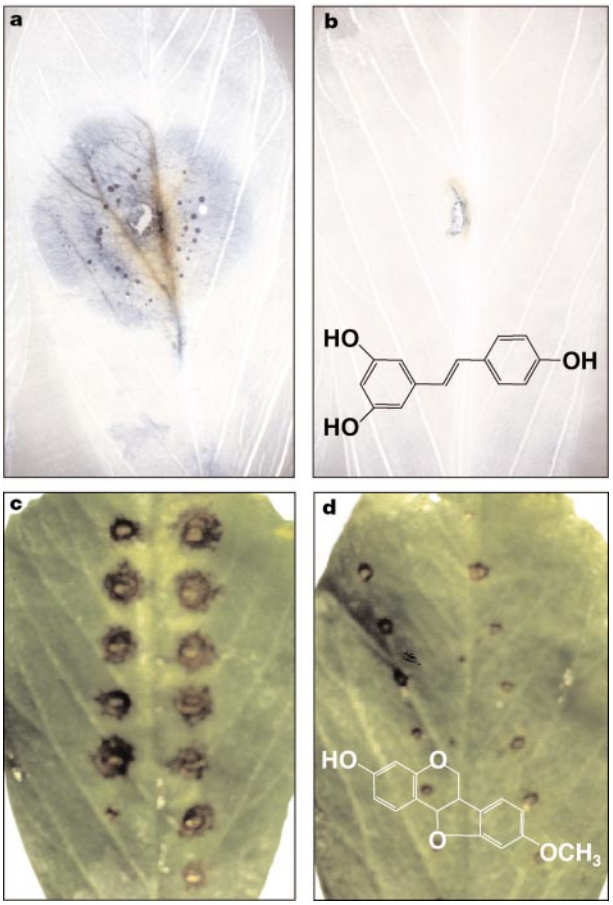


Figure 2 Metabolic engineering for improved fungal disease resistance in alfalfa. **a**, Disease lesion on a leaf of an empty-vector control alfalfa plant 10 days after infection with *Phoma medicaginis* (fungal hyphae and reproductive structures (pycnidia) are stained blue against the background of the leaf that has been cleared of chlorophyll). **b**, Lesion of identical age on a plant that had been transformed with a grapevine stilbene-synthase gene under control of the cauliflower mosaic virus 35S promoter. Such plants constitutively accumulate a glucoside of the novel phytoalexin resveratrol (structure indicated) and exhibit vastly reduced fungal development (data of N. L. Paiva, Noble Foundation[21]). **c**, Lesions of *Phoma medicaginis* 5 days post-inoculation on leaves of alfalfa plants transformed with an empty-vector control construct. **d**, Lesions of identical age on plants constitutively expressing an alfalfa *IOMT* transgene from the 35S promoter. The endogenous *IOMT* genes, and the other genes of the isoflavonoid pathway, are not constitutively expressed, but are induced after fungal infection. The high expression of IOMT during the early stages of the response to the fungus results in more rapid production of the endogenous phytoalexin medicarpin (structure indicated) and consequent reduction in disease severity[22].

impaired in production of phytoanticipins or phytoalexins. Thus, the maize mutation *bxi* abolishes DIMBOA (2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one) biosynthesis, and *bxi* homozygotes are extremely susceptible to northern corn leaf blight (*Helminthosporium turcicum*) and stalk rot (*Diplodia maydis*)[15]. Likewise, saponin-deficient (*sad*) mutants of oats are compromised in resistance to *G. graminis* var. *tritici*, and avenacin content and disease resistance correlate in segregating progeny[16]. This last result confirms the critical role of avenacin in disease resistance in oats, first revealed by disruption of the fungal avenacin detoxifying system.

Several phytoalexin-deficient (*pad*) mutants of *Arabidopsis thaliana* have reduced levels of the indole camalexin (Fig. 1, **21**)[17–19]. The *PAD4* gene seems to encode a regulatory factor[18], whereas *PAD3* encodes a cytochrome P450 that might be directly involved in camalexin biosynthesis[19]. The disease phenotypes of various single

© 2001 Macmillan Magazines Ltd

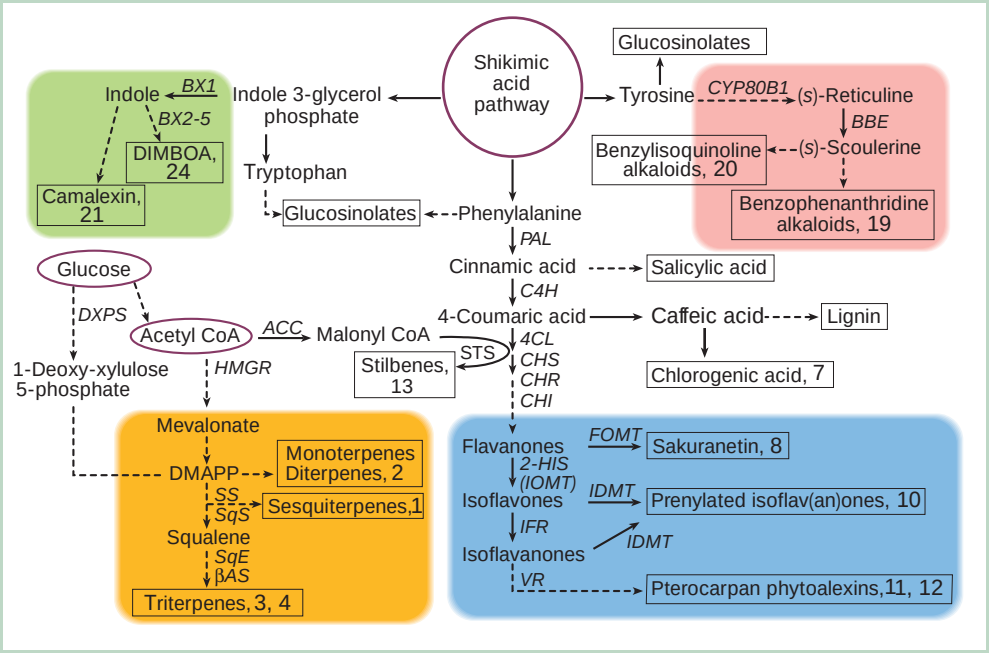


Figure 3 Biosynthetic relationships between antimicrobial plant natural products. Primary precursors/pathways are circled, and end products are boxed. Important cloned genes or characterized enzymes are in italic type. Numbers in bold refer to chemical structures in Fig. 1. Pathways leading to alkaloids are in pink, flavonoids/isoflavonoids in blue, indole derivatives in green, and isoprenoids in orange. BBE, berberine bridge enzyme; PAL, L-phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate CoA ligase; CHS, chalcone synthase; CHR, chalcone (polyketide) reductase; CHI, chalcone isomerase; STS, stilbene synthase; FOMT, flavanone 7-*O*-methyltransferase; 2-HIS, 2-hydroxyisoflavanone synthase ('isoflavone synthase'); IOMT, isoflavone 4′-*O*-methyltransferase; IDMT, isoflavone or isoflavanone dimethylallyl transferase; IFR, isoflavone reductase; VR, vestitone reductase; BX1, indole-3-glycerol phosphate lyase; BX2-5, four consecutive cytochrome P450 enzymes of DIMBOA biosynthesis; DXPS, 1-deoxy-xylulose 5-phosphate synthase; ACC, acetyl CoA carboxylase; HMGR, 3-hydroxy-3-methylglutaryl CoA reductase; SS, sesquiterpene synthase; SqS, squalene synthase; SqE, squalene epoxidase; βAS, β-amyrin synthase. Note that salicylic acid is a signal molecule for plant defence responses (see article in this issue by Stuiver and Custers, pages 865–868).

and double *pad* mutants reveal a complex relationship between phytoalexin production and disease resistance that is highly pathogen-dependent. Thus, the *pad-3* mutant shows increased susceptibility to two fungal pathogens, *Cochliobolus carbonum* and *Alternaria brassicola*, but not to *Botrytis cinerea*, or to the obligate pathogens *Peronospora parasitica* (downy mildew) or *Erysiphe orontii* (powdery mildew)[18,20]. However, in spite of a total lack of camalexin production, *pad-3* does not show increased susceptibility to strains of the bacterial pathogen *Pseudomonas syringae*[17]. In contrast, *pad-1* and *pad-2* show increased susceptibility to disease-causing strains of *P. syringae*, but not to strains to which the wild-type *Arabidopsis* plants were resistant[17,18]. This suggests that camalexin production may limit disease symptoms, but is not responsible for limiting ingress of the pathogen in resistant interactions.

Targeted transgenic approaches allow for evaluation of effects of directly altered phytoalexin profiles, and such approaches have been undertaken in the case of stilbenoids and isoflavonoids. Introduction of a novel phytoalexin, resveratrol (Fig. 1, **13**), into alfalfa by constitutive expression of a grapevine stilbene-synthase gene resulted in reduced symptoms following infection by the leaf spot pathogen *Phoma medicaginis*[21] (Fig. 2a, b). Constitutive overexpression of isoflavone *O*-methyltransferase (IOMT) in transgenic alfalfa resulted in more rapid and increased production of the pterocarpan phytoalexin medicarpin (Fig. 1, **12**) after infection by *P. medicaginis*, resulting in amelioration of symptoms[22] (Fig. 2c, d).

Taken together, the results of genetic analyses indicate that phytoalexins and phytoanticipins can indeed be effective in contributing to resistance *in vivo*, but more than one individual compound or class of compound may be necessary to impart resistance, which is consistent with the multi-component nature of plant defence responses. Clearly, the diversity of plant natural products and host–pathogen combinations makes it impossible to make any general conclusions in regard to the multitude of systems not yet analysed.

Enzymology and organization of biosynthetic pathways

The diversity and apparent complexity of biosynthetic pathways of plant natural products was for many years seen as a barrier to progress in advancing the understanding of phytoalexin function, and in developing technologies to improve resistance by pathway engineering. However, there have recently been a number of successful applications of genetic and genomic approaches to identify genes of plant natural product biosynthesis. For example, comparative EST database mining was used to identify the 2-HIS cytochrome P450 that catalyses the entry-point reaction into isoflavonoid phytoalexin biosynthesis[23,24], based on the predicted expression of this enzyme for production of isoflavonoids in developing seeds and elicited cell cultures (Fig. 3). Mass sequencing of complementary DNA libraries corresponding to metabolically specialized cells has been used to identify several of the genes of monoterpene biosynthesis and the associated Rhomer pathway for formation of the isoprenoid precursor isopentenyl diphosphate in mint glandular trichomes[25,26]. Tagging with the *Mutator* transposable element was used to map and subsequently clone the first enzyme specific for DIMBOA (Fig. 1, **24**) biosynthesis in maize[15] (Fig. 3). Virtually all the enzymes of flavonoid and isoflavonoid biosynthesis[27,28], the early steps of mono-, sesqui- and diterpene biosynthesis[29], and the formation of DIMBOA[15] have now been isolated and functionally characterized. Figure 3 presents an overview of these and other pathways leading to phytoalexins or phytoanticipins, indicating how they are inter-related through primary metabolism, and highlighting some of the genes that have been cloned so far.

Although information derived from large-scale genomics initiatives is invaluable, it can be potentially difficult to mine these data to construct whole pathways from gene sequence information alone. With few exceptions (for example, the BX1 indole-3-glycerol phosphate lyase and four subsequent cytochrome P450 enzymes (BX2–BX5) of DIMBOA biosynthesis[15]; Fig. 3), genes encoding enzymes in natural product pathways are not closely linked in plants, making genetic characterization of pathways more difficult than in bacteria or fungi. However, bioinformatic tools are available for identifying genes encoding members of specific classes of enzymes, such as cytochrome P450s[30], *O*-methyltransferases (OMTs)[31], terpene cyclases[29] and polyketide synthases[32] from large EST projects. In addition, knowledge of transcript expression patterns (for example, upregulated after pathogen attack) from DNA-microarray or more conventional membrane-hybridization approaches can sufficiently

© 2001 Macmillan Magazines Ltd

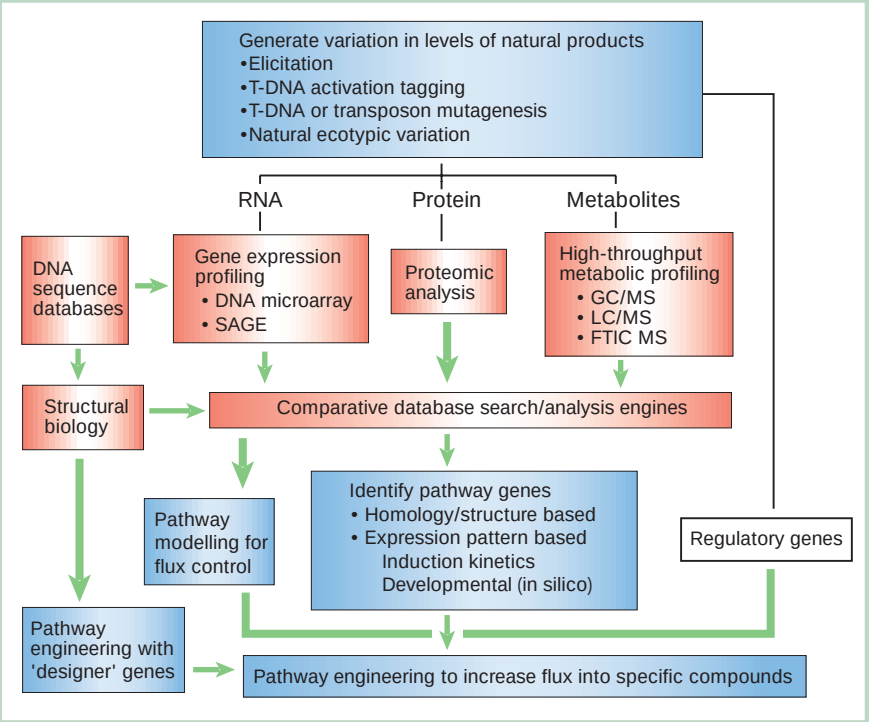


Figure 4 A genomics approach to understanding and manipulating complex natural product pathways for plant defence. A current impediment to progress is a lack of understanding of the biosynthetic and regulatory genes of many natural product pathways. Requirements for a genomics-based approach are large DNA sequence (genomic or EST) databases and ability to experimentally manipulate the levels of the metabolites of interest. This can be done by chemical elicitation, which removes the complication of the presence of the infecting organism, by T-DNA activation tagging to upregulate regulatory genes or rate-limiting enzymes of the pathway, or by exploiting variation in metabolite levels or composition among natural ecotypes or cultivar germplasm. The key to this type of analysis will be the development of improved computational methods for comparative analysis of databases housing information on relative changes in metabolites, proteins and gene transcripts, thereby facilitating the linking of sets of genes or gene products to particular metabolic pathways. SAGE, serial analysis of gene expression; GC/MS and LC/MS, gas chromatography or liquid chromatography linked to mass spectrometry; FTIC MS, Fourier transform ion cyclotron mass spectrometry.

reduce the number of candidate EST clones to allow functional identification by heterologous expression in bacteria, yeast or insect cells.

Sequence-based gene annotation for enzymes in natural product pathways can occasionally be misleading because of the evolutionary flexibility of plant secondary metabolism. For example, chalcone synthase is a simple homodimeric polyketide synthase (PKS) catalysing the first committed step in flavonoid biosynthesis using 4-coumaroyl coenzyme A as the starter molecule (Fig. 3). Many genes annotated as chalcone synthase probably encode different PKS enzymes that use different starter molecules or different numbers of malonyl CoA additions to generate different types of polyketide derivatives[32]. Two acyl transferases of natural product biosynthesis have recently been shown to be closely related to serine proteases, including possession of the catalytic triad diagnostic for this class of protease[33,34], and would have been mistaken for such in the absence of full functional identification.

The methyltransferase (IOMT, Fig. 3) that introduces the 4′-methoxyl function into isoflavonoids *in vivo* methylates the 7-position of isoflavones *in vitro*[22]. In this case, and in other pathways leading to antimicrobial natural products[35,36], consecutive enzymes are physically associated in complexes through which intermediates are channelled, impacting both the kinetics and the regiospecificity of the reactions. Understanding the molecular architecture of such complexes will have a large impact on our ability to engineer new metabolic pathways for crop protection. At a higher order of subcellular organization, some phytoalexins may be synthesized, or at least accumulate, in specialized vesicles that are then delivered to the site of microbial infection, as in the case of the red deoxyanthocyanidin phytoalexins of sorghum[37].

Metabolic engineering for plant disease resistance

The concept of improving disease resistance by engineering natural product pathways has met with several objections, despite a generally held belief that many crop plants are susceptible to pathogens because of years of selective breeding leading to removal of natural products found in their more resistant, wild counterparts (see review in this issue by Stuiver and Custers, pages 865–868). The first objection concerns the large numbers of genes that may have to be transferred, and coordinately regulated, to introduce effective antimicrobial activity. This still represents a technological challenge, although several single-step conversions can generate antimicrobial compounds from ubiquitous or common metabolic intermediates (for example, *O*-methylation of the flavanone naringenin to yield sakuranetin, isoprenylation of isoflavones, or production of stilbenes and other polyketides from malonyl CoA and various starter molecules, as shown in Fig. 3). Increasing production of an endogenous antimicrobial compound through overexpression of a rate-limiting enzyme is conceptually a simple strategy, but in most cases the flux control points in the pathway are not understood. Transferred DNA (T-DNA) activation tagging has recently been applied to the characterization of transcriptional regulators of natural product pathways[38,39], and this approach, by which regulatory genes can be identified without the need to understand the individual enzymatic steps of the pathway, offers an exciting opportunity to develop new molecular tools for pathway engineering for improved plant defence.

A second problem is the ability of pathogens to overcome effects of antimicrobial compounds by rapidly evolving detoxification systems, often involving cytochrome P450 enzymes[40]. It would now seem possible to circumvent this problem by introduction of more than one unrelated, new antimicrobial compound, using the approaches outlined above. However, such a strategy remains to be tested.

Recently, significant progress has been made in elucidating the three-dimensional structures of several key enzymes involved in the biosynthesis of terpenoid and isoflavonoid phytoalexins[41–43]. These studies open up the possibility of structure-based rational design of enzymes such as terpene cyclases, polyketide synthases and *O*-methyltransferases for transgenic introduction of new natural products for plant defence. For example, alfalfa chalcone synthase has been converted into a pyrone synthase by rational introduction of six point mutations[44], and the alterations in plant OMT substrate specificities that result in some cases from heterodimer formation[45] could form a basis for 'combinatorial biochemistry' approaches to pathway engineering

Future prospects

The concept of phytoalexins as induced antimicrobial compounds in plants was first developed in 1940[46]. It has taken over 60 years to gain a basic understanding of how a small part of nature's plant antimicro-

© 2001 Macmillan Magazines Ltd

bial arsenal is manufactured. Much more rapid progress can now be expected. To use genomics to identify genes of plant natural product biosynthesis, it is necessary to be able to perturb the production of these products, whether they be naturally constitutive or inducible. This can be achieved by elicitation with plant or microbial signal molecules[47,48], or by T-DNA activation tagging. The latter method has the advantage of direct identification of regulatory genes or rate-limiting enzymes for the pathway under consideration[49]. Recent improvements in high-throughput metabolite profiling, using gas chromatography or liquid chromatography linked to mass spectrometry, now make it possible to screen for changes in the levels of several hundred plant metabolites in a single sample[50,51]. Linking changes in metabolite profiles to parallel analysis of gene expression through DNA microarray analysis[52], and of protein levels through mass spectrometric proteomic analysis[53], provides possibilities for assigning gene functions based on cellular dynamics. This will also facilitate modelling of pathway flux for determination of rate control points for subsequent metabolic engineering.

The conceptual framework for a genomics approach to natural product pathways is illustrated in Fig. 4. Current technical limitations are the throughput rates for metabolite separation and determination of protein structure, the sensitivity of protein separation methods for proteomic analysis[54], and the need to develop algorithms for comparisons between the very different types of databases that store gene, protein and metabolite information. Fourier transform ion cyclotron mass spectrometry removes the requirement for chromatographic separations and may revolutionize proteomic[54] and metabolic profile analysis. Given the pace of recent developments in genomics and bioinformatics, we should soon be evaluating plants with 'designer' natural product profiles for their adaptation to both biotic and abiotic environmental stresses. □

1. Cooper, R. M. *et al.* Detection and cellular localization of elemental sulphur in disease-resistant genotypes of *Theobroma cacao*. *Nature* **379,** 159–162 (1996).
2. VanEtten, H. D., Matthews, D. E. & Matthews, P. S. Phytoalexin detoxification: importance for pathogenicity and practical implications. *Annu. Rev. Phytopathol.* **27,** 143–164 (1989).
3. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
4. Bell, C. *et al.* The *Medicago* genome initiative: a model legume database. *Nucleic Acids Res.* **29,** 114–117 (2001).
5. Pichersky, E. & Gang, D. R. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends Plant Sci.* **5,** 439–445 (2000).
6. Grayer, R. J. & Harborne, J. B. A survey of antifungal compounds from higher plants. *Phytochemistry* **37,** 19–42 (1994).
7. Harborne, J. B. The comparative biochemistry of phytoalexin induction in plants. *Biochem. System. Ecol.* **27,** 335–367 (1999).
8. Hammerschmidt, R. Phytoalexins: what have we learned after 60 years? *Annu. Rev. Phytopathol.* **37,** 285–306 (1999).
9. Mansfield, J. W. in *Mechanisms of Resistance to Plant Diseases* (eds Slusarenko, A., Fraser, R. S. S. & van Loon, L. C.) 325–370 (Kluwer, Dordrecht, 2000).
10. VanEtten, H., Mansfield, J. W., Bailey, J. A. & Farmer, E. E. Two classes of plant antibiotics: phytoalexins versus "phytoanticipins". *Plant Cell* **6,** 1191–1192 (1994).
11. Kodama, O., Suzuki, T., Miyakawa, J. & Akatsuka, T. Ultraviolet-induced accumulation of phytoalexins in rice leaves. *Agric. Biol. Chem.* **52,** 2469–2473 (1988).
12. Pierce, M. L., Cover, E. C., Richardson, P. E., Scholes, V. E. & Essenberg, M. Adequacy of cellular phytoalexin concentrations in hypersensitively responding cotton leaves. *Physiol. Mol. Plant Pathol.* **48,** 305–324 (1996).
13. Bowyer, P., Clarke, B. R., Lunness, P., Daniels, M. J. & Osbourn, A. E. Host range of a plant pathogenic fungus determined by a saponin detoxifying enzyme. *Science* **267,** 371–374 (1995).
14. Enkerli, J., Bhatt, G. & Covert, S. F. Maackiain detoxification contributes to the virulence of *Nectria haematococca* MP VI on chickpea. *Mol. Plant-Microbe Interact.* **11,** 317–326 (1998).
15. Frey, M. *et al.* Analysis of a chemical plant defense mechanism in grasses. *Science* **277,** 696–699 (1997).
16. Papadopoulou, K., Melton, R. E., Leggett, M., Daniels, M. J. & Osbourn, A. E. Compromised disease resistance in saponin-deficient plants. *Proc. Natl Acad. Sci. USA* **96,** 12923–12928 (1999).
17. Glazebrook, J. & Ausubel, F. M. Isolation of phytoalexin-deficient mutants of *Arabidopsis thaliana* and characterization of their interactions with bacterial pathogens. *Proc. Natl Acad. Sci. USA* **91,** 8955–8959 (1994).
18. Glazebrook, J. *et al.* Phytoalexin-deficient mutants of *Arabidopsis* reveal that *PAD4* encodes a regulatory factor and that four *PAD* genes contribute to downy mildew resistance. *Genetics* **146,** 381–392 (1997).
19. Zhou, N., Tootle, T. L. & Glazebrook, J. Arabidopsis *pad3*, a gene required for camalexin biosynthesis, encodes a putative cytochrome P450 monooxygenase. *Plant Cell* **11,** 2419–2428 (1999).
20. Thomma, B. P. H. J., Nelissen, I., Eggermont, K. & Broekaert, W. F. Deficiency in phytoalexin production causes enhanced susceptibility of *Arabidopsis thaliana* to the fungus *Alternaria brassicola*. *Plant J.* **19,** 163–171 (1999).
21. Hipskind, J. D. & Paiva, N. L. Constitutive accumulation of a resveratrol-glucoside in transgenic alfalfa increases resistance to *Phoma medicaginis*. *Mol. Plant-Microbe Interact.* **13,** 551–562 (2000).
22. He, X.-Z. & Dixon, R. A. Genetic manipulation of isoflavone 7-*O*-methyltransferase enhances the biosynthesis of 4′-*O*-methylated isoflavonoid phytoalexins and disease resistance in alfalfa. *Plant Cell* **12,** 1689–1702 (2000).
23. Steele, C. L., Gijzen, M., Qutob, D. & Dixon, R. A. Molecular characterization of the enzyme catalyzing the aryl migration reaction of isoflavonoid biosynthesis in soybean. *Arch. Biochem. Biophys.* **367,** 147–150 (1999).
24. Jung, W. *et al.* Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. *Nature Biotechnol.* **18,** 208–212 (2000).
25. Crock, J., Wildung, M. & Croteau, R. Isolation and bacterial expression of a sesquiterpene synthase cDNA clone from peppermint (*Mentha* x *piperita*, L.) that produces the aphid alarm pheromone (*E*)-β-farnesene. *Proc. Natl Acad. Sci. USA* **94,** 12833–12838 (1997).
26. Lange, B. M., Wildung, M. R., McCaskill, D. & Croteau, R. A novel family of transketolases that directs isoprenoid biosynthesis via a mevalonate-independent pathway. *Proc. Natl Acad. Sci. USA* **95,** 2100–2104 (1998).
27. Holton, T. A. & Cornish, E. C. Genetics and biochemistry of anthocyanin biosynthesis. *Plant Cell* **7,** 1071–1083 (1995).
28. Dixon, R. A., Harrison, M. J. & Paiva, N. L. The isoflavonoid phytoalexin pathway: from enzymes to genes to transcription factors. *Physiol. Plant.* **93,** 385–392 (1995).
29. Bohlmann, J., Meyer-Gauen, G. & Croteau, R. Plant terpenoid synthases: molecular biology and phylogenetic analysis. *Proc. Natl Acad. Sci. USA* **95,** 4126–4133 (1998).
30. Chapple, C. Molecular-genetic analysis of plant cytochrome P450-dependent monooxygenases. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49,** 311–343 (1998).
31. Ibrahim, R. K., Bruneau, A. & Bantignies, B. Plant *O*-methyltransferases: molecular analysis, common signature and classification. *Plant Mol. Biol.* **36,** 1–10 (1998).
32. Schröder, J. A family of plant-specific polyketide synthases: facts and predictions. *Trends Plant Sci.* **2,** 373–378 (1997).
33. Li, A. X. & Steffens, J. C. An acyltransferase catalyzing the formation of diacylglucose is a serine carboxypeptidase-like protein. *Proc. Natl Acad. Sci. USA* **97,** 6902–6907 (2000).
34. Lehfeldt, C. *et al.* Cloning of the SNG1 gene of arabidopsis reveals a role for a serine carboxypeptidase-like protein as an acyltransferase in secondary metabolism. *Plant Cell* **12,** 1295–1306 (2000).
35. Hrazdina, G. & Jensen, R. A. Spatial organization of enzymes in plant metabolic pathways. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43,** 241–267 (1992).
36. Winkel-Shirley, B. Evidence for enzyme complexes in the phenylpropanoid and flavonoid pathways. *Physiol. Plant.* **107,** 142–149 (1999).
37. Snyder, B. A. & Nicholson, R. L. Synthesis of phytoalexins in sorghum as a site-specific response to fungal ingress. *Science* **248,** 1637–1639 (1990).
38. van der Fits, L. & Memelink, J. ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science* **289,** 295–297 (2000).
39. Borewitz, J., Xia, Y., Blount, J. W., Dixon, R. A. & Lamb, C. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. *Plant Cell* **12,** 2383–2393 (2001).
40. Maloney, A. P. & VanEtten, H. D. A gene from the fungal plant pathogen *Nectria haematococca* that encodes the phytoalexin-detoxifying enzyme pisatin demethylase defines a new cytochrome P450 family. *Mol. Gen. Genet.* **243,** 506–514 (1994).
41. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science* **277,** 1815–1820 (1997).
42. Jez, J. M., Bowman, M. E., Dixon, R. A. & Noel, J. P. Structure and mechanism of the evolutionarily unique plant enzyme chalcone isomerase. *Nature Struct. Biol.* **7,** 786–791 (2000).
43. Zubieta, C., Dixon, R. A. & Noel, J. P. Crystal structures of chalcone *O*-methyltransferase and isoflavone *O*-methyltransferase reveal the structural basis for substrate specificity in plant *O*-methyltransferases. *Nature Struct. Biol.* **8,** 271–279 (2001).
44. Jez, J. M. *et al.* Structural control of polyketide formation in plant-specific polyketide synthesis. *Chem. Biol.* **7,** 919–930 (2000).
45. Frick, S., Ounaroon, A. & Kutchan, T. M. Combinatorial biochemistry in plants: the case of *O*-methyltransferases. *Phytochemistry* **56,** 1–4 (2001).
46. Muller, K. O. & Borger, H. Experimentelle Untersuchungen uber die Phytophthora-Resistenz der Kartoffel- zugleich ein Bewitrag zum Problem der "erworbenen Resistenz" im Pflanzenreich. *Arb. Biol. Anst. Reichsanst (Berl)* **23,** 189–231 (1940).
47. Dixon, R. A., Harrison, M. J. & Lamb, C. J. Early events in the activation of plant defense responses. *Annu. Rev. Phytopathol.* **32,** 479–501 (1994).
48. Nurnberger, T. Signal perception in plant pathogen defense. *Cell. Mol. Life Sci.* **55,** 167–182 (1999).
49. Weigel, D. *et al.* Activation tagging in Arabidopsis. *Plant Physiol.* **122,** 1003–1013 (2000).
50. Trethewey, R. N., Krotzky, A. J. & Willmitzer, L. Metabolic profiling: a Rosetta Stone for genomics? *Curr. Opin. Plant Biol.* **2,** 83–85 (1999).
51. Fiehn, O. *et al.* Metabolite profiling for plant functional genomics. *Nature Biotechnol.* **18,** 1157–1161 (2000).
52. Kehoe, D. M., Villand, P. & Somerville, S. DNA microarrays for studies of higher plants and other photosynthetic organisms. *Trends Plant Sci.* **4,** 38–41 (1999).
53. Yates, J. R. III Mass spectrometry and the age of the proteome. *J. Mass. Spectrom.* **33,** 1–19 (1998).
54. Smith, R.D. Probing proteomes—seeing the whole picture. *Nature Biotechnol.* **18,** 1041–1042 (2000).

© 2001 Macmillan Magazines Ltd

Programmed cell death, mitochondria and the plant hypersensitive response

Eric Lam, Naohiro Kato & Michael Lawton

Biotechnology Center for Agriculture and the Environment, Rutgers University, 59 Dudley Road, New Brunswick, New Jersey 08901-8520, USA (e-mail: Lam@aesop.rutgers.edu)

The plant response to attempted infection by microbial pathogens is often accompanied by rapid cell death in and around the initial infection site, a reaction known as the hypersensitive response. This response is associated with restricted pathogen growth and represents a form of programmed cell death (PCD). Recent pharmacological and molecular studies have provided functional evidence for the conservation of some of the basic regulatory mechanisms underlying the response to pathogens and the activation of PCD in animal and plant systems. In animals, the mitochondrion integrates diverse cellular stress signals and initiates the death execution pathway, and studies indicate a similar involvement for mitochondria in regulating PCD in plants. But many of the cell-death regulators that have been characterized in humans, worms and flies are absent from the *Arabidopsis* genome, indicating that plants probably use other regulators to control this process.

In any natural habitat, there is continual conflict between pathogenic microbes and larger, multicellular organisms. Viruses require a suitable host cell whose cellular machinery can be co-opted to help produce more viral particles. For bacteria and fungi, infection of the host provides access to nutrients as well as a site for growth and reproduction. The resistance of some plants to infection by certain pathogenic strains reflects the presence of disease resistance (*R*) genes, which are predicted to encode receptors for pathogen-derived molecules (refs 1, 2, and see review in this issue by Dangl and Jones, pages 826–833). The protein products of *R* genes act at, or near, the beginning of signalling pathways that invoke the defence responses. The absence of *R* genes renders plants susceptible to infection and in severe cases they may die. In other cases, the plant, although infected, may outgrow the pathogen for long enough to complete its life cycle.

Unlike animals, plants lack an immune system that produces specialized cells, such as T cells, that can attack, disable and eliminate pathogens. Instead, plants have apparently opted for more general defence strategies. One such strategy, which has analogies in animals, is the induction of PCD. In plants, this phenomenon forms part of the hypersensitive response (HR) and its classification is based mainly on morphological criteria of the resultant cell-death lesions as well as the functional suppression of pathogen growth[3–5]. The HR occurs at the site of pathogen entry and involves PCD in and around the infection site. It is also accompanied by the induction of plant defence responses that serve to confine the pathogen and protect the plant. Figure 1 shows the necrotic lesions characteristic of the HR associated with the resistance response of tobacco plants to tobacco mosaic virus (TMV). In this particular host–pathogen system, the product of the *N* gene is required for the recognition of TMV replicase protein by tobacco cells[6,7]. Plants lacking this gene fail to mount a HR and instead display widespread chlorosis that is associated with the systemic spread of the virus.

PCD: a conserved response to microbial pathogens

In animal cells, the ability of many viruses to replicate and spread is dependent on the production of inhibitors of apoptosis (a form of PCD defined by specific molecular and cytological features). Examples include the E1B 19K protein of adenovirus, which antagonizes the pro-apoptotic protein Bax[8], and the p35 and inhibitor of apoptosis proteins (IAPs) from baculovirus[9], which act by inhibiting caspases, a family of cysteine proteases that serve as the critical switch for many forms of PCD in animal cells[10,11]. Evidence for the functional significance of PCD as a plant defence response against viruses is apparent from studies in which the *p35* gene of baculovirus was expressed in transgenic tobacco plants that were then exposed to TMV (O. del Pozo & E.L., unpublished data). Normally this strain of the virus induces resistance in tobacco cultivars containing the *N* gene, and the replicating virus is confined to small HR lesions (Fig. 1). However, in the *p35*-expressing transgenic plants, cell death was delayed and the virus was able to spread beyond the initial infection site and systemically infect the tobacco plant. In these plants, although the HR is still invoked upon infection, it occurs too slowly to prevent cell-to-cell viral movement. For the HR to be effective, it must occur rapidly and precede pathogen movement.

Although the HR is a common feature of many resistance reactions, it is not an obligatory component. Some resistance reactions, such as those mediated by the *mlo* gene of barley towards the fungus *Erysiphe graminis* f. sp. *hordei*, precede the induction of a visible HR (reviewed in ref. 12). Under high humidity, the *Cf* genes of tomato confer resistance to particular pathotypes of the fungus *Cladosporium fulvum* without invoking a visible HR[13]. In *Arabidopsis*, the *dnd1* (defence with no death) mutant expresses resistance to pathogens that otherwise induce a HR on wild-type plants, although *dnd1* plants can still express a highly attenuated HR under some circumstances[14]. Viral resistance can also be uncoupled from HR-associated cell death. In the case of the *Rx* gene from potato, virus resistance at the level of suppression of virus replication can be activated in the absence of cell-death induction[15]. These observations indicate that some pathogens can be effectively resisted without recourse to the HR, and it has been suggested that the HR may represent a final stage of the resistance response that is invoked only if a certain cellular threshold of defence stress signals is reached[3]. One of the main outstanding questions is whether the HR is

© 2001 Macmillan Magazines Ltd

Figure 1 Tobacco mosaic virus/*N* gene interaction: a classic hypersensitive response (HR) model system. Tobacco leaves (*Nicotiana tabacum* cv. Samsun *NN*) were either mock treated (left) or inoculated with TMV, U1 strain (right). The picture was taken 3 days post-inoculation. In cultivars lacking the *N* gene, no lesion spots are observed and TMV systemically infects the whole tobacco plant, leading to the chlorotic mosaic phenotype from which the virus takes its name. In *N* gene-containing cultivars, TMV replication leads to formation of HR lesions that efficiently restrict the virus to the inoculated regions. Only in rare cases is systemic spreading of the virus observed, following induction of a HR.

important in the response of plants to infection with bacteria and fungi, or whether it is simply invoked as part of a more generalized response to pathogens, regardless of its particular efficacy.

Although the role of cell death in bacterial and fungal defence is not clear in all cases, it does seem to assist in the retardation of pathogen proliferation within the host in some model systems[5]. For necrotrophic pathogens, however, induction of cell death may be used by the pathogen to aid in their invasion of the plant[16]. This may be akin to the case of *Shigella* and *Salmonella* invasion of the animal host, where invasion proteins produced by these pathogenic bacteria activate macrophage PCD and acute inflammatory responses that aid in their propagation[17]. Better understanding of control of cell death during various types of host–pathogen interaction will be invaluable for our appreciation of the role that this process has in either host resistance or pathogen proliferation.

Role of mitochondria in integration of cell-death signals

Biochemical, molecular and genetic studies in animal systems have pointed to the importance of compartmentalization in general, and the mitochondrion in particular, in the regulation of PCD[10,18]. Although PCD can be activated in mammalian cells independently of the mitochondrion (for example, by activation of cell-surface 'death' receptors, such as CD95 (Apo-1/Fas)), roles of this organelle seem to include interpretation as well as amplification of cellular stress signals emanating from different sub-cellular compartments to regulate cell-death activation. In animals, this intracellular communication may involve the translocation to the mitochondrion of proteinaceous signals derived from the nucleus (for example, the steroid hormone receptor TR3), the endoplasmic reticulum (for example, Bcl-2 and Bax), the plasma membrane (for example, caspase-8, activated by death receptors such as CD95), the cytoskeleton (for example, Bim) and the cytosol (for example, Bax, Bak and Bid) (reviewed in ref. 18). *In vitro* studies have shown that these proteins associate with, and modify the permeability of, the outer mitochondrial membrane (OMM). This is thought to reflect the ability of these proteins to form channels, or to interact with and modify other proteins that form ion-conducting channels (for example, by promoting or inhibiting oligomerization of channel subunits). In addition to responding to proteinaceous intracellular death signals, the mitochondrion may also initiate apoptosis in response to changes in the levels of cellular messengers such as calcium, to changes in cellular pH or to changes in the levels of metabolites that reflect the energy status of the cell (for example, ATP, ADP, NADH, NADPH and creatine phosphate)[19].

Increased permeability of the OMM may lead to the release from the mitochondrion of a number of cell-death activators, inhibitors and inhibitor derepressors, including cytochrome *c*, apoptosis-inducing factor (AIF) and Smac/DIABLO (Fig. 2). Among these proteins, cytochrome *c* has been shown to regulate the activity of the initiator caspase, procaspase-9, by means of the adaptor protein Apaf-1 (ref. 10). Caspase-9 in turn recruits and activates the effector caspases responsible for ordered disassembly of the cell. Smac/DIABLO, on the other hand, activates intracellular caspases indirectly by a derepression mechanism in which cellular IAPs that hold procaspases in check are competed by interaction with the amino terminus of Smac[20]. AIF, another executor of PCD released from the mitochondrion, is a large protein that can enter the nucleus and activate the fragmentation of chromatin DNA to 50-kilobase-pair fragments[21]. It seems likely that the mitochondrion functions as an important storehouse for key cell death-signalling proteins, which are kept physically sequestered from their targets until the appropriate execution order has been given. Disruption of the OMM and leakage of cytochrome *c* to the cytosol may also lead to inhibition of electron flow from complex III to complex IV in the inner mitochondrial membrane (IMM). This in turn may lead to generation of reactive oxygen species (ROS) that can serve as amplification signals for PCD[22].

Several lines of evidence point to the importance of the mitochondrion in the expression of HR-associated PCD in plants. At present, it is not clear whether cytochrome *c* leakage also occurs during the HR, although leakage has been observed in plant cells undergoing PCD in response to other (non-pathogenic) inducers (reviewed in ref. 23). The HR-inducing bacterial virulence factor harpin disrupts mitochondrial functions[24], and HR-like cell death and disease-resistance marker gene expression can also be activated in plant cells in which Bax is expressed from a viral vector[25]. Activation of plant cell death by Bax requires targeting of the expressed Bax protein to the mitochondrion, whereas mutations that affect its oligomerization suppress this activity. These observations suggest that Bax activates cell death in plants using mechanisms that are fundamentally similar to those present in animal cells.

Further evidence for the involvement of plant mitochondria in the regulation of HR-associated cell death comes from studies of the alternative oxidase (AOX), an IMM enzyme that is not found in animal mitochondria. AOX catalyses electron flow directly from ubiquinol to oxygen, thereby creating an electron shunt that bypasses complexes III and IV of the IMM and results in a cyanide-insensitive electron-transfer pathway. AOX activation by treatment with cyanide during the HR may help to suppress cell death during the propagative phase of lesion formation and thus restrict the size of the necrotic zone[26]. Antisense suppression of AOX resulted in hypersensitivity to antimycin A, an inhibitor of complex III[27]. Upon treatment with antimycin A, these transgenic plants undergo rapid cell death, concomitant with the marked production of ROS in the mitochondria and the induction of genes characteristic of the disease-resistance response. Overexpression of AOX has the reverse effects, providing supporting evidence for a model in which plant mitochondria have an important role as a signal generator for HR-induced cell death, perhaps by generation of ROS derived from electron-transfer intermediates in the IMM[23]. These features indicate that AOX may act as a safety valve for the control of ROS generation from the mitochondrion and are consistent with its activation during the

© 2001 Macmillan Magazines Ltd

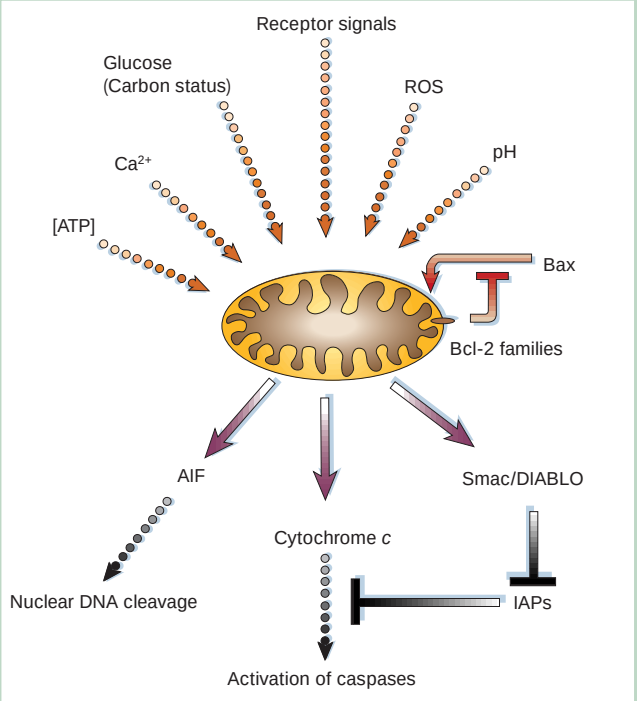


Figure 2 The death signal pathway through the mitochondrion. The model incorporates recent results, mostly from animal model systems, that point to the mitochondrion as a central player in the reception and amplification of cell-death signals. Red lines show the input signals originating from various subcellular sources. Purple lines show the output signals that activate processes leading to programmed cell death. Black lines show secondary signals that may serve to amplify cell-death signals produced by the mitochondrion. Solid lines indicate the signalling components that are associated physically with the mitochondrion. Dotted lines indicate signal and response pathways that may not occur directly. ROS, reactive oxygen species.

HR[28]. This raises the interesting question of how animal cells might regulate ROS production, given the absence of AOX in these cells.

In plants, the mitochondrion is not the only compartment in which ROS can be generated. In addition to a plasma membrane-localized NADH oxidase, which may be analogous to the well-characterized mammalian homologues that mediate activation of the inflammatory response[29,30], the plastid organelles can also participate in HR-associated cell-death signalling. Recently, the plastid-localized protein DS9 was found to regulate the rate of HR cell death in the *N* gene/TMV system[31]. The level of DS9 was suppressed during the onset of the HR, concomitant with a decrease in photosynthetic electron flow. Overexpression of DS9 led to a delay of cell-death activation by TMV, along with increased systemic movement of the virus, even though markers of the HR were still induced. Conversely, suppression of DS9 led to hyperactivation of HR cell death, resulting in smaller lesions that are indicative of more efficient restriction of the virus. Because DS9 encodes a homologue of the bacterial protease FtsH, which is involved in quality maintenance of cellular proteins, these observations indicate that accumulation of damaged proteins in the plastids might act as a signal for HR induction.

Conserved mechanism and mediators

Given the differences in cellular architecture and their evolutionary divergence, it is surprising to find that numerous mediators of disease-resistance signalling in plants share conserved motifs with proteins that are known to have similar roles in the defence response of animals. One of the early surprises resulting from the molecular characterization of signalling components of plant defence was the finding that one class of *R* genes that is found in several different plant species contains an N-terminal domain that resembles the Toll/interleukin-1 receptor (TIR) of *Drosophila* and mammalian systems[32]. In addition, the flanking regions of many of these plant receptor-like proteins contain a nucleotide-binding domain and conserved amino acids, termed the Ap-ATPase (for apoptosis-ATPase) domain[33,34]. This motif is strikingly similar to that present in the mammalian PCD mediator Apaf-1. Downstream targets of the *R* genes include kinases and ankyrin domain-containing proteins that may act to orchestrate changes in gene expression by modulating the activity of a number of transcription factors[32].

In plants, the ankyrin domain-containing mediator NPR1 is essential for activation of systemic acquired resistance. Possible functional equivalents of NPR1 in animal systems include Cactus from *Drosophila* and IκB of mammalian systems. Both Cactus and IκB regulate defence-related gene expression during host–pathogen interactions by interacting with transcription factors that contain the Rel homology domain (for example, NF-κB and Dorsal). But there is no evidence in plants for NPR1 interacting with an NF-κB/Dorsal homologue. Instead, NPR1 interacts with a family of basic-leucine zipper (bZip) domain transcription factors that are conserved in various plant species[35–37]. Differential screening and functional genomic studies have also implicated several other types of transcription factors (containing myb, HD-Zip and WRKY domains) in the HR or disease-resistance response[38–41]. In addition to transcription factors, studies in rice have revealed the role of small GTP-binding proteins of the Rac class as positive regulators of HR-associated cell death and resistance activation for fungal pathogens[42]. GTP-binding proteins, such as Ras, are similarly involved in PCD during eye development in *Drosophila*[43]. These observations indicate that fundamental strategies for responding to microbial invaders and PCD activation may be conserved between animal and plant kingdoms.

The most well-characterized regulatory switch that activates apoptotic cell death in metazoans is a family of specialized cysteine proteases called caspases[10]. Although direct homologues of mammalian caspases are not present in plants, their possible involvement in the activation of HR cell death was inferred from studies in which the HR was attenuated by synthetic peptide inhibitors of caspase activity[44]. Caspase-like protease activity has been demonstrated directly in plants expressing the HR, and in PCD associated with other responses[44,45]. A family of caspase-related proteases (the metacaspases) has recently been identified in plants on the basis of iterative homology searches[11]. The metacaspases identified in *Arabidopsis* fall into two types. Type I contains a predicted caspase-like proteolytic domain that includes the invariant histidine and cysteine residues. This domain is fused to a novel carboxy-terminal domain of ~200 amino acids. Type I caspases lack the death-effector domain or caspase-activating recruitment domain found in the traditional caspase-1 and caspase-3 families. Type II metacaspases contain an N-terminal Zn-Pro domain, which consists of a zinc-finger and a proline-rich repeat motif. This Zn-Pro domain is also found in LSD-1, a protein possibly involved in the control of PCD during the propagation phase of the plant HR[46]. It remains to be seen whether the metacaspases are functionally equivalent to classical caspases in terms of their target specificities as well as their involvement in controlling cell-death activation.

In addition to the caspase superfamily, both plant and animal cells may also use other classes of proteins as effector or initiator proteases to regulate the onset of PCD. In recent years, members of the cathepsin family of proteases have been found to be important in different model systems of developmental and disease-related PCD[47]. For HR cell death in plants, cystatin-sensitive proteases have been found to be critical regulators in a soybean model system[48]. However, the complete repertoire and network of proteases that orchestrate the demise of animal and plant cells remains to be elucidated.

Bax-induced PCD in yeast and plants

In animal cells, mitochondria-mediated PCD acts through the pro-apoptotic Bax family of proteins, which associate with the OMM and oligomerize to form an ion-conducting channel through which

© 2001 Macmillan Magazines Ltd

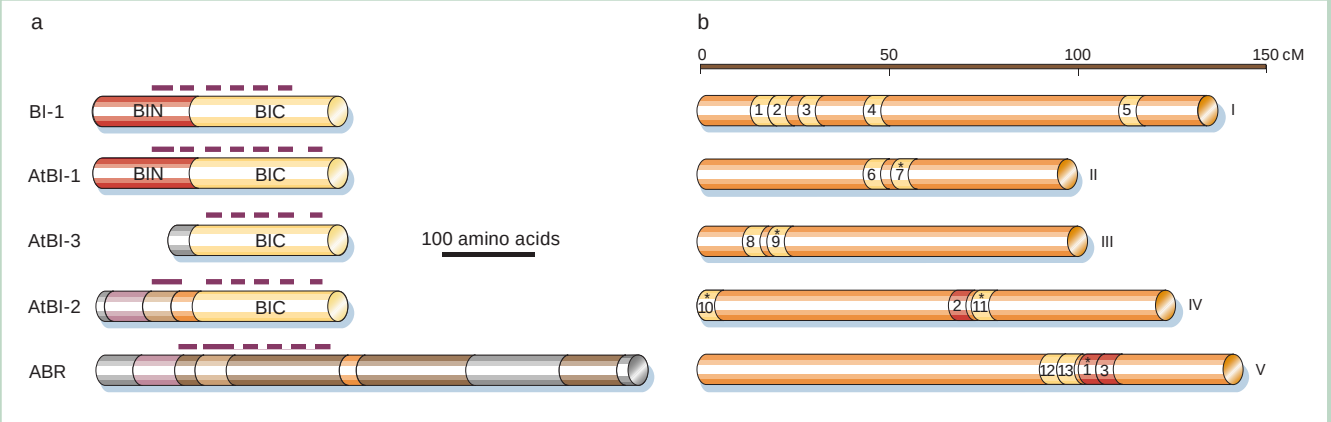


Figure 3 Structural analysis of BI-1-related proteins in the *Arabidopsis* genome. **a**, Predicted domain structures of BI-1-related proteins in *Arabidopsis*. We performed a BLAST search of the GenPept, PIR and SwissPROT databases for plant homologues to human BI-1 (shown at the top of the figure; *Homo sapiens* GI:1729891). A detailed description of the methodology and alignment of the amino-acid sequences of these proteins and their predicted sequences can be found in our website (www.cook.rutgers.edu/~lamlab/ABR.html). The figure shows an alignment between the mammalian BI-1 and several homologues present in *Arabidopsis*. We have divided BI-1 and its homologues into a less well-conserved N-terminal domain (BIN, sharing 42% of the consensus sequence) and a more highly conserved C-terminal domain (BIC, sharing 76% of the consensus sequence). The presence of the BIC domain alone in three of the plant homologues suggests that this division is meaningful and that the BIN and BIC domains may perform different functions. Aside from the sequence homologies, AtBI-1 (GI:8978079) has also been shown to be an inhibitor of mammalian Bax-mediated cell death in yeast, and a rice homologue OsBI-1 functions in a similar yeast assay[53,54]. AtBI-3 (GI:8978080) represents a truncated form of BI-1, and contains only the BIC domain fused to a short N-terminal region that is unrelated to BI-1 or to any other protein sequence in the database. AtBI-2 (also called Ar-BI-1 by Xu and Reed[52]; GI:4455800), which was the first plant BI-1 homologue identified, consists of a novel N-terminal domain fused to a BIC domain. A subsequent search for other sequences that are homologous to the AtBI-2 N-terminal domain revealed a new gene family of 13 members, which all contain a dispersed version of these sequences (indicated by the pink, beige and orange sections in the figure). These predicted proteins (GI: 2252851, 2894606, 4914368, 5306273, 5734718, 6041829, 6664308, 6664309, 6714301, 8885601, 8885602, 10092315, 11994107), which we designated as ABR1-13 (for AtBI-2-related proteins, 1–13), are 40% homologous over a region of 648 amino acids. But the most extensive region of homology occurs in the predicted N-terminal domains and coincides with six predicted hydrophobic, membrane-spanning sequences (shown as short horizontal bars on top). **b**, Location of AtBI- and ABR-encoding genes in the *Arabidopsis* genome. The five chromosomes of the *Arabidopsis* genome are shown as orange bars with the relative size indicated on top in centimorgans (cM). Roman numerals indicate chromosome numbers. Yellow sections indicate positions for ABRs (AtBI-2-related proteins). Red sections indicate AtBI loci. The numbers indicate AtBI and ABR numbers that the different genes have been assigned (see above), and the asterisks on the top of the numbers indicate the genes for which the corresponding cDNAs are found in the public expressed sequence tag collections.

macromolecules and other metabolites can pass[18]. This activity is blocked by the anti-apoptotic proteins Bcl-2 and Bcl-x_L; in contrast, association with members of the Bcl-2 family that contain only the Bcl-2 homology domain 3 (BH3) activates PCD. Although Bcl-2 and Bax are at the heart of mitochondrion-mediated PCD in mammals, comparative genomics reveals that these proteins (and many other regulators of mammalian PCD) are absent from yeast and from plants[34,49]. Nonetheless, expression of Bax in budding yeast[50] and plants[25] causes cell death, and several lines of evidence indicate that this may occur through mechanisms similar to those operating in mammalian cells. Addition of Bax to isolated yeast mitochondria leads to release of cytochrome *c*, while the co-expression of Bcl-2 in yeast cells blocks the cytotoxic effects of Bax[50]. In tobacco, delay of HR-associated cell death as well as repression of irradiation-induced PCD by Bcl-x_L or Ced-9 overexpression were reported[51]. Thus, the expression of pro- and anti-apoptotic Bcl-2-related proteins in these heterologous systems has similar effects to those observed in animal cells. This is consistent with the conjecture that they may regulate a conserved cellular pathway for cell-death control.

Reed and co-workers exploited this functional conservation to screen a human complementary DNA library for inhibitors of Bax-induced PCD in yeast. From this screen, they isolated a new protein, called BI-1 (for Bax inhibitor-1), which was able to inhibit the toxicity of Bax *in vivo*[52]. At present, the mechanism of action of BI-1 is unclear. Localization studies showed that most BI-1 is associated with the heavy membrane fraction (containing mitochondria, lysosomes and rough endoplasmic reticulum) and nuclear envelope. BI-1 can associate with Bcl-2 *in vivo* (through its BH4 domain) but apparently not with Bax, whose activity it antagonizes *in vivo*. Because BI-1 contains six potential transmembrane helices, it has been proposed that it (like Bcl-2 and Bax) may form ion-conducting channels, or modify the activity of existing channels formed by Bcl-2 or Bax family members. However, because yeast does not contain a Bcl-2 homologue, BI-1 must ameliorate the effects of Bax in this system by interacting with, and modulating the properties of, other macromolecules that regulate PCD. Homologues of BI-1 isolated from *Arabidopsis* and from rice have also been shown to suppress Bax-induced PCD in yeast[53,54]. Although direct evidence is absent, BI-1 may function as an anti-apoptotic protein in plants. If so, then by analogy with the yeast studies, it may interact with or affect the activity of proteins that are structurally unrelated but functionally equivalent to mammalian Bcl-2 and Bax proteins.

Although database searches have failed to identify any Bcl-2 homology domain-containing open reading frames in the *Arabidopsis* genome, a search using human BI-1 identified a number of related plant sequences. Alignment of these sequences reveals the domain organization of the predicted *Arabidopsis* BI-1 homologues and related proteins (Fig. 3). AtBI-1 is most highly related to the BI-1 proteins found in animals and contains a highly conserved C-terminal domain and a less well conserved N-terminal domain. This protein was able to suppress Bax-mediated cell death in yeast, and is structurally similar to the rice gene *OSBH-1*, which is also active in the yeast functional assay[53,54]. *Arabidopsis* also contains two other BI-1 homologues: *AtBI-2* (first described by Xu and Reed[52] as *Ar-BI-1*) and *AtBI-3*, identified from more recent genomic sequence analysis. Both of these genes encode proteins that contain the conserved BI-1 C-terminal domain fused to N-terminal sequences of varying length and composition. A subsequent search of the databases with the N-terminal domains of AtBI-2 revealed that these sequences are also present in a new gene family that encodes proteins with the structural hallmarks of transmembrane proteins that can form macromolecular channels. The predicted proteins, which we call ABRs (for AtBI-2-related

© 2001 Macmillan Magazines Ltd

proteins) are ~650 amino acids in length and, intriguingly, contain five or six predicted membrane-spanning helices. However, most of the predicted amino-acid sequence is unique for this family (shown as the brown regions in Fig. 3). Because human BH-1 interacts with Bcl-2 and affects the biological activity of Bax *in vivo*, presumably by altering its ability to oligomerize and form transmembrane channels, it is tempting to speculate that this shared homology region might allow ABR proteins to interact with plant BI-1 proteins, and that these genes might represent functional equivalents of the mammalian Bcl-2 family. It will be interesting to see if any of these genes have pro- or anti-apoptotic activity when introduced into yeast, human or plant cells.

Roles for ion-conducting channels in PCD

Several strands of evidence support the notion that membrane channel-related proteins are involved intimately in mediating HR cell death. The barley *mlo* gene involved in resistance to powdery mildew fungus contains seven transmembrane helices and is located in the plasma membrane with its N terminus exposed extracellularly[55]. Loss of *mlo* results in spontaneous HR-like cell-death lesions that appear in older leaves, which suggests that *Mlo* may normally function as a negative regulator of cell death during the HR. Database searches revealed that there are over 35 *Mlo* family members present in the *Arabidopsis* genome[12], suggesting that these are conserved proteins that may have diverse functional roles within plant cells. The *Arabidopsis* gene *Dnd1*, which is required for the activation of HR cell death by certain host–pathogen combinations, encodes a cyclic nucleotide-dependent calcium channel[14]. This observation is intriguing, in the light of a recent report that cGMP synthesis is required for nitric oxide-induced cell death in *Arabidopsis* and is sensitive to caspase inhibitors[56]. Lastly, expression of the proton channel from *Halobacterium halobium*, bacterioopsin (bO), has been found to activate HR cell death as well as disease resistance in different plant species[57]. Recent site-directed mutational studies as well as epitope-tagging-assisted fractionation showed that bO is expressed in the plasma membrane of plant cells and that passive proton flux through this structurally well-characterized channel is required for its *in planta* activities (D. Pontier, R. Mittler and E.L., unpublished data). These observations suggest that alteration of ionic homeostasis in the cytosol may be crucial in the signalling events that lead to HR cell death.

A recent functional screen for plant proteins that can suppress the death-inducing activity of Bax in yeast has led to the cloning of a tomato glutathione *S*-transferase–peroxidase (GST/GPX)[58]. Expression of this tomato protein restored normal cellular levels of glutathione and preserved the membrane potential across the mitochondrial membrane, both of which had been disrupted by the expression of Bax. In addition, expression of this GST/GPX enzyme in yeast enhanced resistance to oxidative stresses involving ROS. This result suggests that Bax-induced cell death in yeast is mediated through ROS-dependent cellular events that originate from the mitochondrion. Similar observations have been made in animal systems with Bax-mediated cell death and Bcl-x_L-mediated survival[59,60].

Future perspectives

The induction of cell death in plants seems to be a common response to many different types of biotic and abiotic stress. Cell death associated with the HR may be only one of a larger set of cellular responses that are coordinately activated by different stress signals. Understanding the functional role of PCD will require further work. In terms of combating pathogen invasion, it is most likely involved in restricting viral pathogens that have replicated beyond a certain titre within the plant cell. The timing of this cell-death induction relative to the rate of replication of the virus is apparently critical for the effective prevention of virus escape into the phloem of the plant host, after which point systemic infection will ensue. Cell death may also be involved in restricting the growth of obligate bacterial and fungal pathogens that infect living plant tissues. However, the role of the HR in limiting the growth of other types of pathogens is not so clear.

To activate the systematic dismantling of plant cells, mechanisms similar to those revealed in recent years for animal PCD may have been conserved, at least in terms of the general cellular strategies used. For the initial signalling of PCD, the mitochondrion is likely to be important as an integrator of many signals that reflect ionic homeostasis and the metabolic status of the cell. This view is consistent with the observation that many proteins that have now been shown to modulate the threshold for cell-death activation in plants and animals are membrane proteins that can regulate the flux of macromolecules as well as simple ions either at the level of the OMM or the plasma membrane. One of the future challenges will be to create a conceptual framework in which the balance of energy status (for example, ATP and NADH/NADPH levels), ionic conditions (for example, $[Ca^{2+}]$ and pH) and metabolite concentrations (for example, glucose concentration) affect how the mitochondrion is poised to activate cell death by releasing pro-cell-death proteins (for example, cytochrome *c* and AIF). The study of channels present on the OMM, and perhaps other membrane compartments such as the endoplasmic reticulum, will also be important areas in elucidating how cell-death signals, such as ROS or metabolic status, may be translated into a cascade of changes in the permeability of these membranes to macromolecules. In this regard, an efficient reverse genetics approach such as post-transcriptional gene silencing or RNA interference strategies[61], coupled with informatic approaches that can identify a large number of candidate regulators in a completely sequenced genome such as *Arabidopsis*[41,49], may help to speed up the first essential step of identifying the important players involved in cell-death activation. This approach would be complementary to more classical genetic approaches that would reveal new regulators that may not have known homologues[1].

Heterologous systems have opened the door on functional studies of the mechanisms of cell death that operate in plants. One of the main challenges in the next few years will be to understand how different cellular stress signals leading to PCD are perceived, integrated and acted upon in plant cells. It will be particularly interesting to see how the mechanisms that regulate PCD reflect the different subcellular compartments present in plant cells and which must communicate with each other. Although we have focused here on the role of the mitochondrion in PCD, it will be important to determine whether plants, like other eukaryotes, possess cell-death pathways that are mediated independently of the mitochondrion, and whether PCD in plants is actively suppressed by intercellular 'survival signals', as it is in animals[62]. The implication of cell-surface arabinogalactan glycoproteins in the regulation of PCD in *Arabidopsis* cells is particularly intriguing in this regard[63]. □

1. Feys, B. J. & Parker, J. E. Interplay of signaling pathways in plant disease resistance. *Trends Genet.* **16,** 449–455 (2000).
2. Baker, B., Zambryski, P., Staskawicz, B. & Dinesh-Kumer, S. P. Signaling in plant-microbe interactions. *Science* **276,** 726–733 (1997).
3. Morel, J.-B. & Dangl, J. The hypersensitive response and the induction of cell death in plants. *Cell Death Differ.* **4,** 671–683 (1997).
4. Heath, M. C. Nonhost resistance and nonspecific plant defenses. *Curr. Opin. Plant Biol.* **3,** 315–319 (2000).
5. Heath, M. C. Hypersensitive response-related death. *Plant Mol. Biol.* **44,** 321–334 (2000).
6. Dinesh-Kumar, S. P., Tham, W.-H. & Baker, B. Structure-function analysis of the tobacco mosaic virus resistance gene *N*. *Proc. Natl Acad. Sci. USA* **97,** 14789–14794 (2000).
7. Erickson, F. L. *et al.* The helicase domain of the TMV replicase proteins induces the N-mediated defence response in tobacco. *Plant J.* **18,** 67–75 (1999).
8. White, E. Life, death, and the pursuit of apoptosis. *Genes Dev.* **10,** 1–15 (1996).
9. Clem, R. J., Hardwick, J. M. & Miller, L. K. Anti-apoptotic genes of baculoviruses. *Cell Death Differ.* **3,** 9–16 (1996).
10. Green, D. R. Apoptotic pathways: paper wraps stone blunts scissors. *Cell* **102,** 1–4 (2000).
11. Uren, A. G. *et al.* Identification of paracaspases and metacaspases: two ancient families of caspase-like proteins, one of which plays a key role in MALT lymphoma. *Mol. Cell* **6,** 961–967 (2000).
12. Shirasu, K. & Schulze-Lefert, P. Regulators of cell death in disease resistance. *Plant Mol. Biol.* **44,** 371–385 (2000).
13. Hammond-Kosack, K. E., Silverman, P., Raskin, I. & Jones, J. D. G. Race-specific elicitors of *Cladosporium fulvum* induce changes in cell morphology and the synthesis of ethylene and salicylic acid in tomato plants carrying the corresponding *Cf* disease resistance gene. *Plant Physiol.* **110,** 1381–1394 (1996).
14. Clough, S. J. *et al.* The Arabidopsis dnd1 "defense, no death" gene encodes a mutated cyclic nucleotide-gated ion channel. *Proc. Natl. Acad. Sci. USA* **97,** 9323–9328 (2000).

© 2001 Macmillan Magazines Ltd

15. Bendahmane, A., Kanyuka, K. & Baulcombe, D. The Rx gene from potato controls separate virus resistance and cell death responses. *Plant Cell* **11,** 781–791 (1999).
16. Govrin, E. M. & Levine, A. The hypersensitive response facilitates plant infection by the necrotrophic pathogen *Botrytis cinerea*. *Curr. Biol.* **10,** 751–757 (2000).
17. Navarre, W. W. & Zychlinsky, A. Pathogen-induced apoptosis of macrophages: a common end for different pathogenic strategies. *Cell Microbiol.* **2,** 265–273 (2000).
18. Ferri, K. F. & Kroemer, G. Mitochondria—the suicide organelles. *BioEssays* **23,** 111–115 (2001).
19. Vander Heiden, M. G. *et al.* Outer mitochondrial membrane permeability can regulate coupled respiration and cell survival. *Proc. Natl Acad. Sci. USA* **97,** 4666–4671 (2000).
20. Chai, J. *et al.* Structural and biochemical basis of apoptotic activation by Smac/DIABLO. *Nature* **406,** 855–862 (2000).
21. Susin, S. A. *et al.* Molecular characterization of mitochondrial apoptosis-inducing factor. *Nature* **397,** 441–446 (1999).
22. Blackstone, N. W. & Green, D. R. The evolution of a mechanism of cell suicide. *BioEssays* **21,** 84–88 (1999).
23. Lam, E., Pontier, D. & del Pozo, O. Die and let live—programmed cell death in plants. *Curr. Opin. Plant Biol.* **2,** 502–507 (1999).
24. Xie, Z. & Chen, Z. Harpin-induced hypersensitive cell death is associated with altered mitochondrial functions in tobacco cells. *Mol. Plant-Microbe Interact.* **13,** 183–190 (2000).
25. Lacomme, C. & Santa Cruz, S. Bax-induced cell death in tobacco is similar to the hypersensitive response. *Proc. Natl Acad. Sci. USA* **96,** 7956–7961 (1999).
26. Chivasa, S. & Carr, J. Cyanide restores *N* gene-mediated resistance to tobacco mosaic virus in transgenic tobacco expressing salicylic acid hydroxylase. *Plant Cell* **10,** 1489–1498 (1998).
27. Maxwell, D. P., Wang, Y. & McIntosh, L. The alternative oxidase lowers mitochondrial reactive oxygen production in plant cells. *Proc. Natl Acad. Sci. USA* **96,** 8271–8276 (1999).
28. Simons, B. H., Millenaar, F. F., Mulder, L., Van Loon, L. C. & Lambers, H. Enhanced expression and activation of the alternative oxidase during infection of *Arabidopsis* with *Pseudomonas syringae* pv. *tomato*. *Plant Physiol.* **20,** 529–538 (1999).
29. Keller, T. *et al.* A plant homolog of the neutrophil NADPH oxidase gp91^{phox} subunit gene encodes a plasma membrane protein with Ca^{2+} binding motifs. *Plant Cell* **10,** 255–266 (1998).
30. Torres, M. A. *et al.* Six Arabidopsis homologues of the human respiratory burst oxidase (gp91^{phox}). *Plant J.* **14,** 365–370 (1998).
31. Seo, S. *et al.* Reduced levels of chloroplast FtsH protein in tobacco mosaic virus-infected tobacco leaves accelerate the hypersensitive response. *Plant Cell* **12,** 917–932 (2000).
32. Wilson, I., Vogel, J. & Somerville, S. Signalling pathways: a common theme in plants and animals? *Curr. Biol.* **7,** R175–R178 (1997).
33. van der Biezen, E. A. & Jones, J. D. G. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr. Biol.* **8,** R226–R227 (1998).
34. Aravind, L., Dixit, V. M. & Koonin, E. V. The domains of death: evolution of the apoptosis machinery. *Trends Biochem. Sci.* **24,** 47–53 (1999).
35. Zhang, Y., Fan, W., Kinkema, M., Li, X. & Dong, X. Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the *PR-1* gene. *Proc. Natl Acad. Sci. USA* **96,** 6523–6528 (1999).
36. Zhou, J.-M. *et al.* NPR1 differentially interacts with members of the TGA/OBF family of transcription factors that bind an element of the *PR-1* gene required for induction by salicylic acid. *Mol. Plant-Microbe Interact.* **13,** 191–219 (2000).
37. Depres, D., DeLong, C., Glaze, S., Liu, E. & Fobert, P. R. The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. *Plant Cell* **12,** 279–290 (2000).
38. Daniel, X., Lacomme, C., Morel, J. B. & Roby, D. A novel *myb* oncogene homologue in *Arabidopsis thaliana* related to hypersensitive cell death. *Plant J.* **20,** 57–66 (1999).
39. Mayda, E., Tornero, P., Conejero, V. & Vera, P. A tomato homeobox gene (HD-zip) is involved in limiting the spread of programmed cell death. *Plant J.* **20,** 591–600 (1999).
40. Euglem, T., Rushton, P. J., Robatzek, S. & Somssich, I. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* **5,** 199–206 (2000).
41. Maleck, K. *et al.* The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nature Genet.* **26,** 403–410 (2000).
42. Kawasaki, T. *et al.* The small GTP-binding protein Rac is a regulator of cell death in plants. *Proc. Natl Acad. Sci. USA* **96,** 10922–10926 (1999).
43. Bergmann, A., Agapite, J., McCall, K. & Steller, H. The Drosophila gene *hid* is a direct molecular target of Ras-dependent survival signaling. *Cell* **95,** 331–341 (1998).
44. del Pozo, O. & Lam, E. Caspases and programmed cell death in the hypersensitive response of plants to pathogens. *Curr. Biol.* **8,** 1129–1132 (1998).
45. Korthout, H. A. A. J., Berecki, G., Bruin, W., van Duijn, B. & Wang, M. The presence and subcellular localization of caspase 3-like proteinases in plant cells. *FEBS Lett.* **475,** 139–144 (2000).
46. Jabs, T., Dietrich, R. A. & Dangl, J. L. Initiation of runaway cell death in an *Arabidopsis* mutant by extracellular superoxide. *Science* **273,** 1853–1856 (1996).
47. Johnson, D. E. Noncaspase proteases in apoptosis. *Leukemia* **9,** 1695–1703 (2000).
48. Solomon, M., Belenghi, B., Delledonne, M., Menachem, E. & Levine, A. The involvement of cysteine proteases and protease inhibitor genes in the regulation of programmed cell death in plants. *Plant Cell* **11,** 431–444 (1999).
49. Aravind, L., Dixit, V. M. & Koonin, E. V. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. *Science* **291,** 1279–1284 (2001).
50. Zha, H. *et al.* Structure-function comparisons of the proapoptotic protein Bax in yeast and mammalian cells. *Mol. Cell. Biol.* **16,** 6494–6508 (1996).
51. Mitsuhara, I., Malik, K. A., Miura, M. & Ohashi, Y. Animal cell-death suppressors Bcl-x_L and Ced-9 inhibit cell death in tobacco plants. *Curr. Biol.* **9,** 775–778 (1999).
52. Xu, Q. & Reed, J. C. Bax inhibitor-1, a mammalian apoptosis suppressor identified by functional screening in yeast. *Mol. Cell* **1,** 337–346 (1998).
53. Kawai, M., Pan, L., Reed., J. C. & Uchimiya, H. Evolutionally conserved plant homologues of the Bax Inhibitor-1 (BH-1) gene capable of suppressing Bax-induced cell death in yeast. *FEBS Lett.* **464,** 143–147 (1999).
54. Sanchez, P., de Torres Zabela, M. & Grant, M. AtBI-1, a plant homologue of Bax inhibitor-1, suppresses Bax-induced cell death in yeast and is rapidly upregulated during wounding and pathogen challenge. *Plant J.* **21,** 393–399 (2000).
55. Devoto, A. *et al.* Topology, subcellular localization and sequence diversity of the Mlo family in plants. *J. Biol. Chem.* **274,** 34993–35004 (1999).
56. Clarke, A., Desikan, R., Hurst, R. D, Hancock, J. T. & Neill, S. J. NO way back: nitric oxide and programmed cell death in *Arabidopsis thaliana* suspension cultures. *Plant J.* **24,** 667–677 (2000).
57. Mittler, R. Shulaev, V. & Lam, E. Coordinated activation of programmed cell death and defense mechanisms in transgenic tobacco plants expressing a bacterial proton pump. *Plant Cell* **7,** 29–42 (1995).
58. Kampranis, S. C. *et al.* A novel plant glutathione *S*-transferase:peroxidase suppresses Bax lethality in yeast. *J. Biol. Chem.* **275,** 29207–29216 (2000).
59. Harris, M. H., Vander Heiden, M. G., Kron, S. J. & Thompson, C. B. Role of oxidative phosphorylation in Bax toxicity. *Mol. Cell. Biol.* **20,** 3590–3596 (2000).
60. Gottlieb, E., Vander Heiden, M. G. & Thompson, C. B. Bcl-x_L prevents the initial decrease in mitochondrial membrane potential and subsequent reactive oxygen species production during tumor necrosis factor alpha-induced apoptosis. *Mol. Cell. Biol.* **20,** 5680–5689 (2000).
61. Smith, N. A. *et al.* Total silencing by intron-spliced hairpin RNAs. *Nature* **407,** 319–320 (2000).
62. McCabe, P. F., Levine, A, Meijer, P.-J., Tapon, N. A. & Pennell, R. I. A programmed cell death pathway activated in carrot cells at low density. *Plant J.* **12,** 267–280 (1997).
63. Gao, M. & Showalter, A. M. Yariv reagent treatment induces programmed cell death in Arabidopsis cell cultures and implicates arabinogalactan protein involvement. *Plant J.* **19,** 321–331 (1999).

Acknowledgements

Research in the Lam and Lawton laboratories is supported by the New Jersey Commission on Science and Technology. Research on plant cell death in the Lam laboratory is supported by competitive grants from the USDA.

© 2001 Macmillan Magazines Ltd

Surface-to-air signals

Edward E. Farmer

Gene Expression Laboratory, Ecology Institute, University of Lausanne, Biology Building, 1015 Lausanne, Switzerland
(e-mail: edwardelliston.farmer@ie-bpv.unil.ch)

Powerful volatile regulators of gene expression, pheromones and other airborne signals are of great interest in biology. Plants are masters of volatile production and release, not just from flowers and fruits, but also from vegetative tissues. The controlled release of bouquets of volatiles from leaves during attack by herbivores helps plants to deter herbivores or attract their predators, but volatiles have other roles in development and in the control of defence gene expression. Some of these roles may include long-distance signalling within and perhaps between plants.

The plant kingdom abounds with natural chemicals, many of which are volatile. These molecules are chemically diverse, being represented by fatty acid derivatives, terpenes, indole and molecules from other chemical families. Ethylene[1], a molecule implicated in both development and defence, was the first gaseous hormone discovered in nature. It now seems likely that new and fundamental insights could emerge from the study of other plant volatiles that can act as signals within the plant, or can be exported selectively, changing the immediate environment of the producer, its neighbours and attackers.

Far from being passive in the face of attack, plants use many remarkable strategies to increase their chances of survival[2]. Some of the most elegant of these strategies involve the release of specific blends of volatiles. In response to herbivores damaging a single leaf, whole plants release a complex array of volatiles. Maize leaves, for example, release a mixture containing several terpenes, including linalool, in response to elicitors present in the regurgitant of beet armyworm larvae[3]. A powerful elicitor of volatile release, volicitin, has been purified from this secretion and identified as a conjugate of 17-hydrolinolenic acid and L-glutamine[4]. The bouquets of volatiles released in response to attack can have several effects that, in the field, have great importance for plant survival[5]. Chief among these is attraction of predators to the feeding herbivore, but other strategies are being discovered. At night, tobacco plants release a characteristic bouquet in response to the feeding larvae of nocturnal moths[6]. The bouquet repels further egg-ladened female moths from ovipositing on the tobacco leaves. These, and many other studies[7,8], underlie the tremendous importance volatiles play in signalling to insects, but can volatiles also affect neighbouring plants?

Volatile signalling within and between plants

A remarkable example of plant-to-plant signalling, probably unrelated to a role in defence, comes from tobacco plants in which dominant ethylene insensitivity has been engineered through introduction of the *Arabidopsis etr1-1* allele[9]. Normally, the leaves of wild-type tobacco plants tend to stop growing as they approach neighbouring tobacco plants — this may stop them wasting energy producing leaves that would be shaded from useful light. But the transgenic tobacco plants lacked normal social behaviour and their leaves grew over and among the leaves of their neighbours. This is a good indication of plant-to-plant signalling under laboratory conditions, but whether ethylene is a social signal in the field is as yet unclear, as is any role for ethylene in interplant defence signalling. A second example of intraspecific signalling in tobacco involves defence and methyl salicylate. This compound is released from wild-type tobacco leaves infected with tobacco mosaic virus (TMV)[10], a virus that causes necrotic lesion formation in the genotype of tobacco used for these laboratory experiments. Methyl salicylate produced by infected plants both increases the resistance of neighbouring uninfected tobacco plants to TMV infection and also induces the expression of the defence gene *PR1* in uninfected plants. Similar to the development of visual necrotic symptoms, release of methyl salicylate is blocked when infected plants are incubated at 32 °C, and commences when the temperature is lowered to 24 °C. This allows elegant control of volatile release in this laboratory system, although so far there is no evidence for a role of methyl salicylate as a plant-to-plant signal in nature.

One of the best studied volatile signals in plants is the fragrant compound methyl jasmonate (MJ), which has been studied as a volatile signal *in planta* and also in laboratory and field experiments in plant-to-plant signalling. The compound is the methyl ester of 3*R*,7*S*-jasmonic acid (Fig. 1), and both are potently active as regulators of gene expression[11,12]. Production of methyl jasmonate can be manipulated *in vivo* using an *Arabidopsis* gene encoding jasmonic acid carboxyl methyltransferase (JMT)[13], which methylates non-volatile jasmonic acid to produce volatile methyl jasmonate. *JMT* has been expressed constitutively in *Arabidopsis* under the control of a powerful viral promoter. The *JMT* transgenic plants had threefold-increased levels of methyl jasmonate compared with wild-type plants, although levels of jasmonic acid were unchanged. The plants were resistant to a fungal pathogen and showed constitutive expression of an inducible defence gene (*PDF1.2*) encoding an antimicrobial peptide. The study raises tractable questions, such as whether or not the conversion of jasmonic acid to methyl jasmonate is a control point *in vivo*. This might be addressed by strategies aimed at lowering *JMT* expression in *Arabidopsis* and other species. It is clear that volatile regulators of gene expression, acting as hormones, can affect the individual that produces them, but it is also possible that atmospheric transport of a signal from its source to distal parts of the same individual could occur (the term 'automone' might be suitable for such compounds).

Interspecific airborne signalling (Fig. 2) involving methyl jasmonate is also known and may relate to defence against wounding organisms. Many and perhaps most plants do not release significant levels of the compound. But it is produced by some flowers, including scented jasmine, as well as in the leaves of several species of *Artemisia*, most

© 2001 Macmillan Magazines Ltd

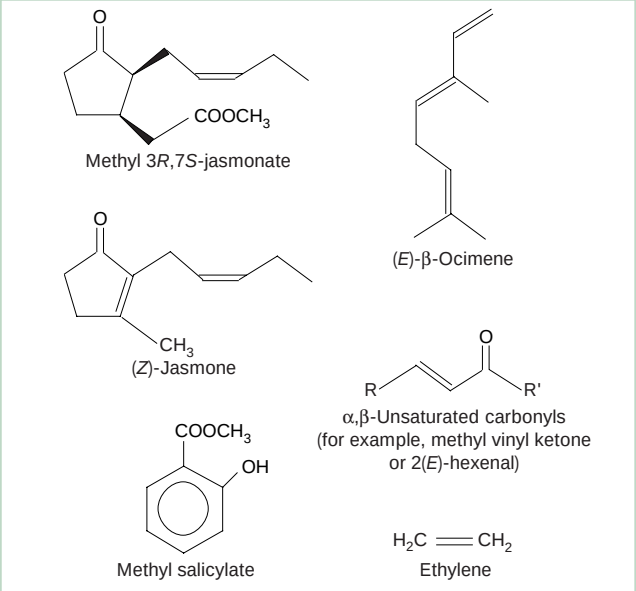


Figure 1 Volatile regulators implicated in plant defence and plant-to-plant information transfer. For α,β-unsaturated carbonyl-containing molecules, R and R′ are small substituents such as protons, methyl groups or alka(e)nyl groups; (*Z*)-jasmone would also fall into this category. In the laboratory, exposure to volatile regulators activates defence gene expression, but very low levels of the compounds, as might occur in nature, could possibly prime or sensitize the defence system of receiver plants, perhaps allowing them to respond faster to future attack.

notably big sagebrush (*A. tridentata*), a plant that dominates large areas of the Great Basin in the western United States. Placing tomato plants and sagebrush branches together for two days in a closed container resulted in the strong accumulation of defence-related proteinase inhibitors in the leaves of tomato. This was found to be due to the release of methyl jasmonate from the sagebrush[14].

This experiment inevitably led researchers to question whether long-distance methyl jasmonate signalling could take place in the field and, if so, what is its significance?[15,16] Karban *et al.*[17] have now addressed the first of these questions. In their experiment, tomato was replaced by a wild tobacco species, *Nicotiana attenuata*, which can grow in the presence of sagebrush in nature. Airborne levels of methyl jasmonate at a distance 3 m from sagebrush plants were found to be below the limit of detection. Next, ~10% of sagebrush leaves on a plant were removed or damaged and a gas trap was fit onto a damaged branch. Within one hour, a 6.5-fold increase of 3*R*,7*S*-methyl jasmonate was registered, proving damage-induced release of a biologically active enantiomer of methyl jasmonate in the field. As a presumed consequence of its release, natural herbivore damage to *N. attenuata* was reduced significantly in tobacco plants growing within 15 cm of the clipped sagebrush. Blocking air contact between sagebrush and tobacco prevented these effects. There is a complementary and perhaps simpler explanation for methyl jasmonate release by sagebrush, a role in plant–plant competition. Because methyl jasmonate is so physiologically active, it is possible that release of the compound could interfere with the growth of neighbouring plants — this aspect deserves much more attention.

Links between plant–insect and plant–plant signalling

Yet more candidate airborne plant-to-plant signals have been reported for bean plants (*Phaseolus lunatus*) infested with herbivorous spider mites (*Tetranychus urticae*). In response to attack, infested leaves release volatiles that can increase the resistance of uninfested leaves to attack by spider mites[18,19], as well as inducing the expression of several defence-related genes in neighbouring uninfected lima bean leaves. β-Ocimene (Fig. 1) and two other related terpenoids are thought to be responsible for this effect[19]. These compounds were shown recently to activate the expression of a number of defence genes in detached bean leaves, but it will be important to test whether the compounds upregulate gene expression in intact bean plants[8]. Potentially related to this work is the interesting biology of (*Z*)-jasmone, a product of jasmonic acid metabolism[20]. (*Z*)-Jasmone released by plants was found to be electrophysiologically active in insects, both herbivores and their predators[21]. Treatment of healthy bean (*Vicia faba*) plants with (*Z*)-jasmone induced both (*E*)-β-ocimene release and also α-tubulin gene expression[21]. One could speculate that, in bean, (*Z*)-jasmone treatment causes release of β-ocimene, which itself activates gene expression, but at present the evidence is lacking. The physiological relevance of α-tubulin gene expression is not yet known, but it is clear that because the expression of other, as yet uncharacterized genes might be affected, the α-tubulin gene serves as a good marker. What is evident from these and other studies is that some compounds, like (*E*)-β-ocimene and (*Z*)-jasmone, can affect both insect behaviour and gene expression in plants, and whenever a molecule is implicated in plant–plant signalling its relevance in plant–insect signalling should be investigated.

Interplant communication: universality or opportunism?

Interplant communication involving airborne regulators of gene expression can occur in the laboratory and might also occur in the field. But it is too early to make generalizations about the frequency of this phenomenon in nature and, as the distances between plants increase, good evidence for plant-to-plant signalling becomes scarcer. Is the phenomenon likely to be universal or restricted to a few examples? An argument against universality concerns chance, opportunity and the rich diversity of molecules made in the plant kingdom. Even considering that the release of volatiles from flowers and leaves is highly controlled and sensitive to many environmental factors[22,23], a few molecules capable of eliciting gene expression will be released, here and there, in the plant kingdom. Therefore, it might be simply a question of putting the right plant species together in the laboratory to see an effect. Furthermore, if intraspecific plant–plant airborne signalling was widespread, there might be problems for the plants concerned. Volatile release might, in some cases, serve to inform congeners of impending attack, but it could be used opportunistically by neighbouring, competing species, which might gain a selective advantage if they could perceive this valuable information. To avoid this, one could imagine selection pressure to develop species-specific volatile signals, a process that could lead to signal diversification in nature. Alternatively, the plant might just cease releasing the volatile in question.

An argument for universality, at least in intraspecific systems, is the number of reports of plant-to-plant signalling where airborne

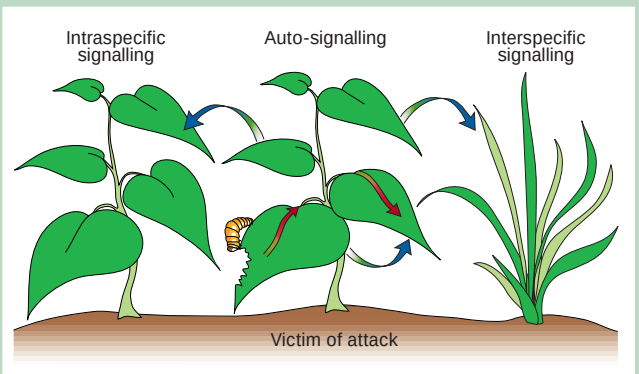


Figure 2 Communicating danger with airborne signals. Four modes of signalling from or within diseased or wounded plants are indicated: signalling to healthy congeners, signalling to members of other species, or auto-signalling either within (arrow in leaf) or outside the plant body. Good evidence exists for plant-to-plant airborne signalling in the laboratory, but field studies are limited.

© 2001 Macmillan Magazines Ltd

signals are implicated but have yet to be characterized[7,24–29]. Recent work on European black alder trees (*Alnus glutinosa*) showed that intact alders close to manually defoliated individuals subsequently showed decreased herbivore damage compared with more distant individuals, with airborne signals from the damaged trees implicated[28]. It will take a sustained effort to bring systems such as these into the molecular arena where the signals produced in the field can be characterized. But it would be worth the effort. Perhaps dense swards of grass would also offer an attractive searching ground for new intra- and interspecific airborne signalling systems?

Volatile electrophiles

Some compounds may have escaped detailed attention with regard to activating gene expression in diseased tissues. These include the electrophile 2(*E*)-hexenal, which is produced by many trees and shrubs, particularly upon wounding, and also as an odour component of various fruit, including cucumber, banana and apple. 2(*E*)-Hexenal is a biocidal molecule that is produced in response to bacterial pathogenesis[30]. Although described many years ago as a widespread volatile antibiotic[31], the broader biological significance of 2(*E*)-hexenal production has received remarkably little attention during the past 20 years. Recently, this compound has been shown to induce the accumulation of sesquiterpenoid phytoalexins in wounded cotton[32] and, in common with a hexenyl acetate isomer(s), induced stress-related gene expression in *Arabidopsis*[33]. 2(*E*)-Hexenal (but not isomers of hexenyl acetate) contains an α,β-unsaturated carbonyl group, and it seems likely that this electrophilic feature in a molecule will confer the ability to induce stress and defence responses in plants[34]. This small reactivity feature alone, presented in volatile form as acrolein, leads to cell damage and to the powerful expression of the glutathione-*S*-transferase 1 (*GST1*) gene in *Arabidopsis*[34]. Electrophiles such as 2(*E*)-hexenal or acrolein are susceptible to nucleophilic attack, for example Michael addition, but it is not known whether they have to undergo chemical reactions in the plant cell in order to produce their effects.

Reactive electrophile species probably are crucial in microbial disease, perhaps even contributing to the damage of cells undergoing hypersensitive (programmed) cell death (ref. 34, and see review in this issue by Lam, Kato and Lawton, pages 848–853). Their volatile counterparts could also be considered in future studies as candidates in information transfer from plant to plant. However, because they are antibiotics, it is possible that the simple absorption of released electrophiles onto the leaves of a healthy plant could lead to increased 'resistance' to microbial pathogens. Some studies on volatile signalling in plants[27] need re-evaluating in this light.

Each of the many facets of research on volatiles — plant–insect signalling, intraplant signalling and plant–plant signalling — could have exciting applications. In the quest for new ways to control insect pests, the chemical induction of volatile release has great practical potential. This has been demonstrated with tomato plants that were induced (by spraying with jasmonic acid) to release a volatile blend attractive to parasitic wasps. Caterpillars feeding on these chemically induced plants were more often victims to the parasites than were caterpillars feeding on control plants[35]. Some volatiles can be remarkably powerful regulators of gene expression in plants and their exploitation in engineered plant defence strategies is now on the horizon, with candidate genes such as *JMT*[13] already in hand. □

1. Bleeker, A. B. & Kende, H. Ethylene: a gaseous signal molecule in plants. *Ann. Rev. Cell Dev. Biol.* **16,** 1–18 (2000).
2. Walling, L. L. The myriad plant responses to herbivores. *J. Plant Growth Reg.* **19,** 195–216 (2000).
3. Turlings, T. C. J & Tumlinson, J. H. Systemic release of chemical signals by herbivore-injured corn. *Proc. Natl Acad. Sci. USA* **89,** 8399–8402 (1992).
4. Alborn, H. T. *et al.* An elicitor of plant volatiles from beet armyworm oral secretion. *Science* **276,** 945–949 (1997).
5. Kessler, A. & Baldwin, I. T. Defensive function of herbivore-induced plant volatile emissions in nature. *Science* **291,** 2141–2144 (2001).
6. De Moraes, C. M., Mescher, M. C. & Tumlinson, J. H. Caterpillar-induced nocturnal plant volatiles repel conspecific females. *Nature* **410,** 577–580 (2001).
7. Dicke, M. & Bruin, J. (eds) Chemical information transfer between wounded and unwounded plants. *Biochem. Syst. Ecol.* **29** (2001).
8. Pickett, J. A. & Poppy, G. M. Switching on plant genes by external chemical signals. *Trends Plant Sci.* **6,** 137–139 (2001).
9. Knoester, M. *et al.* Ethylene-insensitive tobacco lacks nonhost resistance against soil-borne fungi. *Proc. Natl Acad. Sci. USA* **95,** 1933–1937 (1998).
10. Shulaev, V., Silverman, P. & Raskin, I. Airborne signalling by methyl salicylate in plant pathogen resistance. *Nature* **385,** 718–721 (1997).
11. Creelman, R. A. & Mullet, J. E. Biosynthesis and action of jasmonates in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48,** 355–387 (1997).
12. Reymond, P. & Farmer, E. E. Jasmonate and salicylate as global signals for defense gene expression. *Curr. Opin. Plant Biol.* **1,** 404–411 (1998).
13. Seo, H. S. *et al.* Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-regulated plant responses. *Proc. Natl Acad. Sci. USA* **98,** 4788–4793 (2001).
14. Farmer, E. E. & Ryan, C. A. Interplant communication: airborne methyl jasmonate induces synthesis of proteinase inhibitors in plant leaves. *Proc. Natl Acad. Sci. USA* **87,** 7713–7716 (1990).
15. Bruin, J., Sabelis, M. W. & Dicke, M. Do plants tap SOS signals from their infested neighbours? *Trends Ecol. Evol.* **10,** 167–170 (1995).
16. Shonle, I. & Bergelson, J. Interplant communication revisited. *Ecology* **76,** 2660–2663 (1995).
17. Karban, R., Baldwin, I. T., Baxter, K. J., Laue, J. G. & Felton, G. W. Communication between plants: induced resistance in wild tobacco plants following clipping of neighboring sagebrush. *Oecologia* **125,** 66–71 (2000).
18. Bruin, J., Dicke, M. & Sabelis, M. W. Plants are better protected against spider-mites after exposure to volatiles from infested conspecifics. *Experentia* **48,** 525–529 (1992).
19. Arimura, G.-I. *et al.* Herbivory-induced volatiles elicit defense genes in lima bean leaves. *Nature* **406,** 512–515 (2000).
20. Koch, T., Bandemer, K. & Boland, W. Biosynthesis of cis-jasmone: a pathway for the inactivation and the disposal of the plant stress hormone jasmonic acid to the gas phase? *Helv. Chim. Acta* **80,** 838–850 (1997).
21. Birkett, M. A. *et al.* New roles for *cis*-jasmone as an insect semiochemical and in plant defense. *Proc. Natl Acad. Sci. USA* **97,** 9329–9334 (2000).
22. Loreto, F. & Sharkey, T. D. Isoprene emission by plants is affected by transmissible wound signals. *Plant Cell Environ.* **16,** 563–570 (1993).
23. Logan, B. A., Monson, R. K. & Potosnak, M. J. Biochemistry and physiology of foliar isoprene production. *Trends Plant Sci.* **5,** 477–481 (2000).
24. Baldwin, I. T. & Schultz, J. C. Rapid changes in tree leaf chemistry induced by damage: evidence for communication between plants. *Science* **221,** 277–279 (1983).
25. Rhoades, D. F. in *Plant Resistance to Insects* (ed. Hedin, P.) 55–68 (Am Chem. Soc., Washington DC, 1983).
26. Haujioka, E., Suomela, J. & Neuvonen, S. Long-term inducible resistance in birch foliage: triggering cues and efficacy on a defoliator. *Oecologia* **65,** 363–369 (1985).
27. Fujiwara, M., Oku, H. & Shiraishi, T. Involvement of volatile substances in systemic resistance of barley against *Erysiphe graminis* f. sp. *hordei* induced by pruning leaves. *J. Phytopath.* **120,** 81–84 (1987).
28. Dolch, R. & Tscharntke, T. Defoliation of alders (*Alnus glutinosa*) affects herbivory by leaf beetles on undamaged neighbours. *Oecologia* **125,** 504–511 (2000).
29. Rhoades, D. F. Pheromonal communication between plants. *Rec. Adv. Phytochem.* **19,** 195–218 (1985).
30. Croft, K. P. C., Jüttner, F. & Sluzarenko, A. J. Volatile products of the lipoxygenase pathway evolved from *Phaseolus vulgaris* (L.) leaves inoculated with *Pseudomonas syringae* pv *phaseolicola*. *Plant Physiol.* **101,** 13–24 (1993).
31. Lyr, H. & Banasiak, L. Alkenals, volatile defense substances in plants, their properties and activities. *Acta Phytopath. Acad. Sci. Hung.* **18,** 3–12 (1983).
32. Zeringue, H. J. Jr Effects of C_6-C_{10} alkenals and alkanals on eliciting a defense response in the developing cotton boll. *Phytochem.* **31,** 2305–2308 (1992).
33. Bate, N. J. & Rothstein, S. J. C_6-volatiles derived from the lipoxygenase pathway induce a subset of defense-related genes. *Plant J.* **16,** 561–569 (1998).
34. Vollenweider, S., Weber, H., Stolz, S., Chételat, A. & Farmer, E. E. Fatty acid ketodienes and fatty acid ketotrienes: Michael addition acceptors that accumulate in wounded and diseased *Arabidopsis* leaves. *Plant J.* **24,** 467–476 (2000).
35. Thaler, J. S. Jasmonate-inducible plant defences cause increased parasitism of herbivores. *Nature* **399,** 686–688 (1999).

© 2001 Macmillan Magazines Ltd

Co-evolution and plant resistance to natural enemies

Mark D. Rausher

Evolution, Ecology and Organismal Biology Group, Department of Biology, Duke University, Durham, North Carolina 27708-0338, USA (e-mail: mrausher@duke.edu)

Co-evolution between plants and their natural enemies is generally believed to have generated much of the Earth's biological diversity. A process analogous to co-evolution occurs in agricultural systems, in which natural enemies adapt to crop resistance introduced by breeding or genetic engineering. Because of this similarity, the investigation of resistance mechanisms in crops is helping to elucidate the workings of co-evolution in nature, while evolutionary principles, including those derived from investigation of co-evolution in nature, are being applied in the management of resistance in genetically engineered crops.

The process of co-evolution between plants and their natural enemies — including viruses, fungi, bacteria, nematodes, insects and mammals — is believed by many biologists to have generated much of the Earth's biological diversity[1,2]. Co-evolution is a reciprocal evolutionary interaction between a plant and one or more of its natural enemies that occurs in cycles[3–7]. In the first phase of a cycle, natural selection imposed by enemies causes the evolution of a new plant resistance character that reduces enemy attack. Much of the extraordinary chemical and morphological diversity among plant species is believed to reflect this type of defensive adaptation occurring independently in different sets of co-evolving species (Table 1, and see review in this issue by Dixon, pages 843–847). Because most resistance characters reduce the survival or virulence of natural enemies, their evolution generates selection that initiates the second phase of a co-evolutionary cycle: the evolution of counter-resistance by those enemies, that is, the evolution of characters that circumvent the newly evolved plant resistance. Plant natural enemies exhibit a wide variety of physiological, behavioural and morphological characters that seem to have evolved in this way (Table 1).

A process analogous to co-evolution also occurs in agricultural systems. Breeders release a resistant crop variety, and the evolution of counter-resistance typically follows. When breeders respond by introducing another resistant variety, a new cycle is initiated. A typical example of this 'artificial' co-evolution is the attempt to breed wheat resistant to Hessian fly, *Mayetiola destructor*, in Indiana. In 1955, a cultivar carrying a resistance gene was deployed and provided effective resistance. Within six years, however, substantial counter-resistance had evolved in the Hessian fly. A cultivar carrying a second resistance gene was released in 1964, with counter-resistance appearing within eight years. Counter-resistance to a third gene, released in 1971, had evolved within about 10 years[8–10].

Conventional breeding of resistance suffers from a serious limitation: reproductive barriers between species prevent introduction of resistance genes into a crop from any plants except very closely related wild relatives. A potentially effective defence found in, say, teosinte (*Zea mexicana*) can be bred into maize (recently derived from teosinte) but not into rice or soya beans because the latter cannot be crossed with teosinte. The perfection of genetic transformation technology in the 1980s removed this limitation and now allows resistance genes to be transferred into crops from distantly related (even non-plant) species. Such technological advances have stimulated plant molecular biologists to explore the genetic and biochemical control of resistance characters, with the aim of transferring new resistance genes into crops, although they have not alleviated the threat that natural enemies will evolve counter-resistance to whatever resistance genes are deployed.

As I describe below, this threat has stimulated investigations of how evolutionary principles can be applied to slow the evolution of counter-resistance. The molecular analysis of resistance has given renewed impetus to the science of applied evolution, a science that draws upon knowledge gained from the study of natural co-evolution. At the same time, the molecular analysis of resistance has provided new insight into the operation of co-evolution in nature. In particular, it has provided significant evidence for the common operation of the first phase of co-evolutionary cycles, and has facilitated empirical confirmation of one of the main assumptions about how plant defences against enemies evolve.

Natural enemies impose selection for resistance

Evidence indicating that natural enemies generally evolve to overcome the detrimental effects of plant resistance characters (the second phase of co-evolutionary cycles) is abundant: natural enemies exhibit numerous characters that can be interpreted only as having evolved to confer counter-resistance (Table 1). For example, seeds of the tropical legume *Dioclea megacarpa*, which contain the non-protein amino acid L-canavanine, are toxic to most insects because their arginyl-tRNA synthetases also incorporate L-canavanine into proteins. However, the bruchid beetle *Caryedes brasiliensis*, whose larvae feed solely on *D. megacarpa*, has evolved a modified tRNA synthetase that distinguishes between L-canavanine and arginine[11]. Adaptation by natural enemies is also seen at the level of local populations. In parsnip webworms, cytochrome P450 enzymes detoxify furanocoumarins produced by their host plant, wild parsnip (*Pastinaca sativa*). In each population, P450 activity, as well as its specificity towards different furanocoumarins, reflects the concentration and profile of furanocoumarins produced by the host population — each webworm population seems finely adapted to overcoming the particular defences it encounters[12].

By contrast, the contention that apparently defensive traits of plants have actually evolved in response to natural

© 2001 Macmillan Magazines Ltd

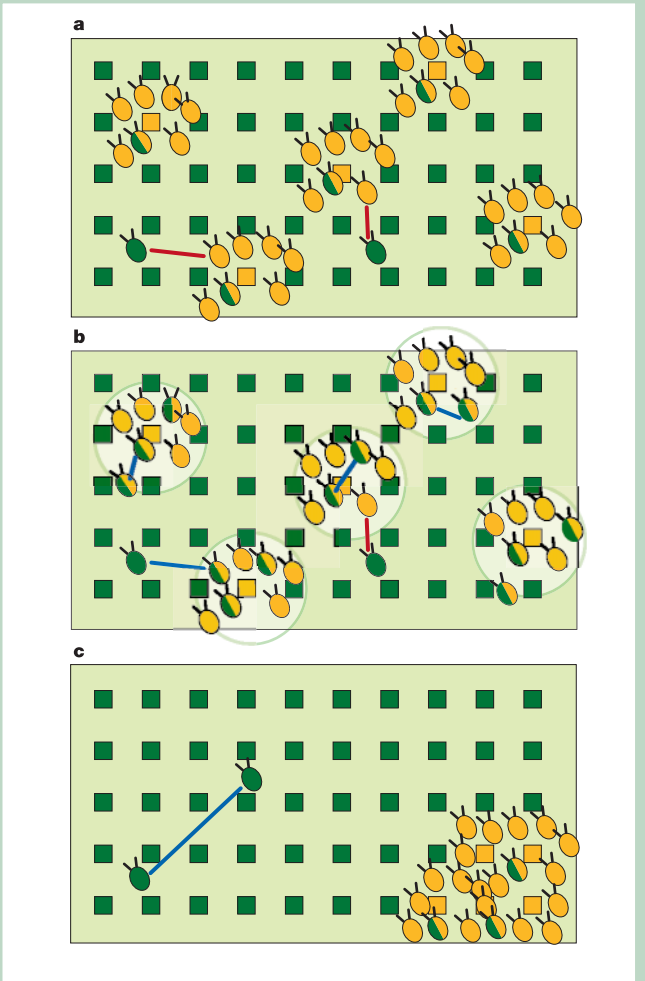


Figure 1 Schematic portrayal of the HDR strategy. Green and yellow squares represent resistant and susceptible (refuge) plants, respectively. Yellow insects, *AA* (susceptible) genotype; yellow and green insects, *Aa* (susceptible) genotype; green insects, *aa* (counter-resistant) genotype. **a**, Refuge plants are interspersed with susceptible plants, and the juvenile insect stage is confined to a single plant. The HDR strategy functions well. The numerous *AA* insects produced on refuge plants screen the few counter-resistant *aa* insects that emerge from resistant plants from each other and ensure that *aa* individuals mate only with *AA* individuals (red lines). This prevents the production of homozygous offspring that would be adapted to feeding on *aa* plants. **b**, Refuge plants are interspersed with susceptible plants, and the juvenile insect stage moves among plants. Circles indicate radius of juvenile insect movement. The HDR strategy functions poorly. Under these conditions, more *Aa* individuals survive because the toxicity of the food plant is diluted. Matings between *Aa* and *aa* or between *aa* and *aa* individuals (blue lines) are thus more frequent and result in the production of *aa* (counter-resistant) offspring. **c**, Refuge plants are clumped, and adult insect dispersal is limited. The HDR strategy functions poorly. Even with movement by juvenile insect stages among plants, most heterozygotes do not feed on susceptible plants and thus die. However, limited dispersal of *AA* insects prevents them from screening matings between the rare *aa* individuals (blue lines), resulting in production of homozygous counter-resistant individuals.

selection imposed by natural enemies is more controversial. For some traits, it is clear that the only function is defence. Thorns and urticating hairs, for example, almost certainly function primarily to protect plants from mammalian herbivores. But in most cases, characters that confer resistance may have additional physiological or ecological functions. For example, although various flavonoids exhibit antifungal and antibacterial properties, most also absorb ultraviolet radiation efficiently and are believed to protect the plant from this environmental hazard[12]. Other functions performed by plant secondary chemicals include conferring frost tolerance, allelopathy, nutrient storage, structural reinforcement, mediation of stigma–pollen interactions, regulation of biochemical processes, and signalling to mutualists[13–17].

As has been repeatedly argued[3,14–16,18–21], the multiplicity of functions attributed to most resistance factors undermines the inference that they have evolved primarily in response to natural enemies. Rather, resistance may simply be a fortuitous side effect of characters that evolved to perform other ecological functions. It has also been argued that in nature, plant enemies are generally too rare to cause the frequent evolution of defensive traits[19,20,22,23]. These arguments call into question the widespread belief that co-evolution between plants and their enemies is common and generates much of the morphological and chemical diversity plants exhibit.

One recent approach to addressing this controversy has been to use manipulative field experiments to compare the pattern of selection on purported defensive characters in the presence and absence of natural enemies[21]. Although the number of cases examined is still small, enemy-imposed selection on resistance has been demonstrated in all investigations conducted, for characters as varied as glucosinolate content and trichome density in *Arabidopsis*[24], alkaloid content in *Datura*[25], resistance to fungal pathogens in *Silene*[21,26], and resistance and tolerance to insects in *Ipomoea*[27,28]. These investigations provide strong evidence for the potential of natural enemies to cause the evolution of plant resistance characters, although they provide little indication of whether resistance characters actually do evolve in response to selection imposed by natural enemies.

An alternative source of information regarding this controversy is being provided as a direct result of investigations of the molecular and biochemical basis of resistance to pathogens. Gene-for-gene resistance seems to be mediated by signal cascades that initiate both localized cell death around the site of pathogen infection, and mobilization of systemic induced resistance[29–31]. At the head of the signal cascade is a receptor protein that recognizes some molecular feature of the invading pathogen (the 'elicitor') and activates the cascade.

The high specificity of the receptor to the pathogen elicitor almost unquestionably indicates that the receptor, and its coupling to the signal cascade, represent adaptations to defence against pathogens. Moreover, these receptors have undergone numerous amino-acid substitutions over a relatively short period of time[32–35]. Because the rate of non-synonymous (amino-acid changing) substitution is often higher than the rate of synonymous (non-amino-acid changing) substitutions in the genes coding for these receptors[33,34], much of this evolutionary change seems to be adaptive rather than due to genetic drift[36]. Finally, amino-acid substitutions are concentrated in regions of the receptor that are believed to interact with the elicitor molecule, as is expected if these regions were co-evolving with those elicitors. These patterns indicate that in the progenitors of crop plants as diverse as rice, tomato, flax and sugar beets, receptor genes conferring resistance have in fact evolved in response to selection imposed by bacterial, fungal or viral pathogens.

Chitinase evolution in *Arabidopsis* and related species in the genus *Arabis* exhibits remarkable similarities to receptor evolution[37]. Plant chitinases are believed to defend against fungal infection by attacking chitin, a principal component of fungal cell walls. Class I chitinases in *Arabis* species often exhibit higher rates of non-synonymous than synonymous substitutions, a hallmark of adaptive evolution. Moreover, amino-acid substitutions are concentrated in the molecule's active site, a pattern not usually seen in enzyme evolution. Because the structure of chitin does not evolve, the most reasonable interpretation of this pattern is that plant chitinases are co-evolving with pathogen chitinase inhibitors, small carbohydrate or protein molecules that competitively inhibit the breakdown of chitin by chitinase.

Because receptor proteins and chitinases represent only a small sample of purported plant defences against natural enemies, it may be premature to generalize from these examples that those traits have generally evolved in response to selection imposed by natural

© 2001 Macmillan Magazines Ltd

enemies. Nevertheless, it is clear that as plant cell and molecular biologists identify the genes associated with other resistance characters, evolutionary biologists will be presented with many more opportunities to address this issue.

Defensive traits are costly

The idea that adaptation is costly is a deeply entrenched principle in evolutionary biology. In the context of plant defences, this principle states that the incremental fitness benefit associated with genotypes conferring increased defence is accompanied by a decrement in fitness associated with reallocation of resources away from other fitness-enhancing functions[38].

For example, an increase in the production of alkaloid compounds as defences against herbivorous insects removes nitrogen from the pool available to a plant for growth, and is thus expected to reduce plant size and seed production. Because such costs are an integral component of the standard evolutionary model for the evolution of resistance[39,40], the validity of this model depends on whether such costs normally exist.

Recent experiments on natural populations of plants as diverse as *Arabidopsis*, *Ipomoea* (morning glories), *Pastinaca* and *Trifolium* (clovers) have provided strong evidence for costs[28,41–45]. These experiments typically use quantitative genetic approaches to determine whether, in the absence of enemies, fitness and resistance are inversely correlated. Their interpretation relies on a crucial assumption: because in the absence of enemies, benefits associated with resistance cannot be realized, any fitness differences among genotypes must reflect pleiotropic effects of resistance genes (that is, multiple effects of single genes, which affect more than one phenotypic character). Although this interpretation is reasonable, it has been difficult to rule out an alternative interpretation: fitness differences in the absence of enemies result from linkage disequilibrium between resistance alleles and alleles at linked loci. If this alternative interpretation were correct, apparent costs of defence could be transient historical effects, rather than permanent constraints on the evolution of defences, as linkage disequilibrium is expected to decay over time.

Distinguishing between these alternative interpretations has been difficult because in most natural systems, little is known about genes associated with resistance. However, in a set of experiments made possible by prior molecular characterization of resistance to the herbicide chlorosulphuron in *Arabidopsis thaliana*[46,47], resistance costs have been shown to result from pleiotropy[48,49]. In *A. thaliana*, chlorosulphuron resistance is conferred by a single base-pair substitution in the gene encoding acetolactate synthase (ALS), which catalyses the first step in the biosynthesis of branched-chain amino acids. By transforming the resistant *ALS* allele into a chlorosulphuron-susceptible *Arabidopsis* stock, isogenic lines that differed only in whether the resistance allele was present were created. In the field, in the absence of herbicides, the resistant lines produced 34% fewer seeds, indicating a substantial fitness cost clearly due to pleiotropy. This cost is believed to be due to either increased metabolic drain caused by overexpression of branched-chain amino acids or a build-up of the toxic α-amino butyric acid.

Although the *ALS* mutant confers resistance to a synthetic herbicide rather than a natural enemy, there is no reason to believe that alleles for resistance to natural enemies would be less likely to have associated pleiotropic costs. Our growing understanding of the molecular action of genes that control resistance to natural enemies should facilitate similar experiments designed to determine whether the apparent costs detected by quantitative genetics experiments are generally due to pleiotropy. Data obtained from such experiments should eventually reveal whether a basic assumption about constraints on resistance evolution in natural populations is valid.

Table 1 Examples of co-evolution in natural plant–enemy systems

Plant defence	Plant taxon	Natural enemy	Counter-resistance	Refs
Toxic furanocoumarins	Umbelliferae	Black swallowtail butterfly	Cytochrome P450 detoxifying enzymes	95, 96
Toxic amino acids	Various Leguminosae	Bruchid weevil	Modified tRNA synthetase	11
Trichomes	*Solanum*	Ithomiid butterfly	Silk scaffolding	97
Latex	*Asclepias* (milkweeds)	Monarch butterfly and others	Leaf-vein-cutting behaviour	98
Enlarged fruits	Sapindales	*Jadera* bugs	Elongated mouthparts	99, 100
Chitinase	*Arabis* (Cruciferae)	Fungal pathogens	Chitinase inhibitors	37
Hypersensitive response *R* genes	Several taxa	Fungal and bacterial pathogens	Modification of elicitor proteins	33–35
Mutualism with predacious ants	*Acacia*	*Polyhymno* (Gelechiid lepidopteran)	Shelter construction	101, 102

Resistance management by evolutionary engineering

Although evolutionary biology is largely an academic science, the practical benefits of the application of evolutionary principles are beginning to be realized in areas as diverse as disease management[50,51], fisheries management[52–54], conservation[55,56], biomolecular engineering[57,58] and computer design[59]. But perhaps it is in the area of resistance management that this potential has begun to be realized most. Resistance management attempts to prevent natural enemies from evolving counter-resistance to pesticides or resistant crops, a hitherto almost inevitable phenomenon. Although the use of evolutionary principles in this area was aimed initially at preserving the effective lifetime of pesticides and conventionally bred crop resistance, it has taken on added importance for genetically engineered crops because of the greatly increased economic investment required for their development and deployment.

Although completely preventing the evolution of counter-resistance may in most cases be impossible, evolutionary biologists have developed strategies that, at least in theory, will slow the evolution of counter-resistance and thus prolong the usefulness of genetically engineered resistance. With these strategies, biologists hope to redirect the course of evolution.

The high-dose/refuge strategy

Recent research has focused on devising strategies for delaying the evolution of counter-resistance by insect herbivores to toxins, particularly *Bacillus thuringiensis* toxins[60]. Although various approaches have been suggested[61,62], researchers, the United States Environmental Protection Agency, and some large corporations are currently concentrating on the 'high-dose/refuge' (HDR) strategy[60,62–65]. This approach involves engineering a crop to produce high doses of a toxin and planting mixes of resistant and susceptible varieties (Fig. 1a).

The functioning of the HDR strategy relies on several basic evolutionary principles. The first principle is that the rate at which an advantageous allele increases in frequency in a population depends greatly on its degree of dominance (Box 1). A completely recessive allele spreads much more slowly than a dominant or additive allele because initially it is present only in heterozygotes, which are sheltered from selection. The 'high-dose' portion of the strategy is aimed at ensuring that alleles conferring counter-resistance are effectively recessive. Even if the LD_{50} (median lethal dose) of the heterozygote is intermediate between that of the homozygotes (that is, counter-resistance is neither dominant nor recessive), a high enough dose of the toxin will still kill more than 99% of the heterozygotes, rendering counter-resistance effectively recessive.

The second evolutionary principle used by the HDR strategy is that the rate of increase in allele frequency is proportional to the difference in fitness between genotypes (Box 1). The 'refuge' portion of the strategy is designed to slow the spread of a counter-resistance allele by reducing the fitness difference between the homozygote for that allele and the other genotypes. The refuge is provided by the non-resistant plants, which allow the susceptible insects to reproduce

© 2001 Macmillan Magazines Ltd

Box 1
Theoretical basis of the HDR strategy

Evolution of resistance conferred by a single mutation can be modelled[74] using the basic population genetic equation for a change in gene frequency at a single locus with two alleles:

$$p' = p(pW_{AA} + qW_{Aa})/(p^2W_{AA} + 2pqW_{Aa} + q^2W_{aa})$$

where p and q are the frequencies of alleles A and a, and W_{ij} is the fitness of genotype ij. Iteration of this equation yields the trajectory of change in gene frequency for allele A[103]. The HDR strategy for resistance management is based on two basic properties of this equation.

1. The rate of increase in the frequency of a new mutant depends on degree of dominance. When a mutant allele first appears, it is rare and occurs only in heterozygotes. The effective magnitude of selection on this allele therefore depends on whether it is expressed in the heterozygote. If the allele is dominant, it is exposed to selection immediately and increases rapidly in frequency, whereas if it is recessive, it is initially shielded from selection and increases only slowly. An example of this effect is portrayed in Fig. B1a, in which the relative fitnesses of the mutant and wild-type alleles are 1 and 0.5 respectively.

2. The rate of increase in the frequency of a new mutant depends on the difference in fitness between genotypes. For the case of a recessive mutant allele, the relevant fitness difference, s, is that between the wild-type and mutant homozygotes. As shown in Fig. B1b, reducing this fitness difference retards the spread of the mutant allele.

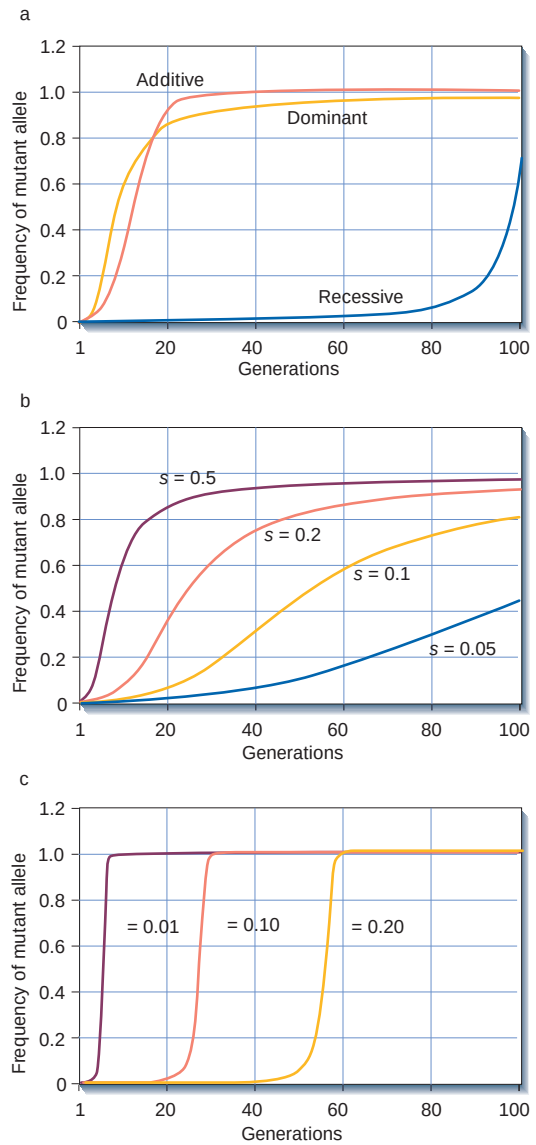
When refuges are used, the fitness of each genotype, W, depends on the proportion of the habitat dedicated to the refuge, symbolized by β (Box 1 Table). With complete recessivity ($h = 0$), the fitness difference between homozygotes is $1 - c - \beta$, indicating that this fitness difference decreases as the size of the refuge increases. For an initial mutant frequency of 0.005, Fig. B1c shows that although counter-resistance evolves to appreciable frequencies within about five pest generations in the absence of a refuge ($\beta = 0.01$), 10% and 20% of refuges delay the development of counter-resistance for more than 25 and 50 generations, respectively.

Box 1 Table Fitness of genotypes *AA*, *Aa* and *aa*

Genotype	Non-refuge	Refuge*	W
AA (susceptible)	0	1	β
Aa (susceptible)	0	$1 - hc$	$\beta(1 - hc)$
aa (resistant)	$1 - c$	$1 - c$	$1 - c$

*c, cost of counter-resistance; h, dominance of heterozygote.

Figure B1 Effects of dominance (**a**), strength of selection (**b**) and refuge size (**c**) on rate of increase of frequency of mutant allele.



substantially. With even a small refuge, the fitness difference is reduced markedly, greatly retarding the rate of increase of the counter-resistance allele (Box 1). In essence, the refuge produces enough susceptible insects to greatly reduce the probability that the rare counter-resistant insects mate with each other and produce counter-resistant offspring (Fig. 1a).

A third evolutionary principle — adaptations are generally costly — suggests that in many situations, the HDR strategy may postpone the evolution of counter-resistance indefinitely. Accumulating evidence indicates that counter-resistance to toxins is frequently costly[66–68], as evidenced by the rapid elimination of counter-resistance when pesticide use is discontinued[69–71]. Such costs, if manifested in heterozygotes, render the overall fitness of heterozygotes less than the fitness of the susceptible homozygote, which will tend to prevent the allele conferring counter-resistance from increasing in frequency when rare. Unfortunately, our ignorance about how often counter-resistance is costly in heterozygotes prevents us from being able to predict how likely genetically engineered resistance is to be evolutionarily stable for indefinite periods under the HDR strategy.

Even in the absence of costs, the theoretical considerations described above indicate that this strategy may greatly slow the evolution of counter-resistance. It is this theoretical promise that has convinced the Environmental Protection Agency and some corporations to adopt the HDR strategy as a goal for resistance management. However, the theoretical predictions of this strategy have been examined empirically only rarely, usually under artificial conditions[72,73], and rest on largely untested assumptions about the basic biology of insect pests. One such assumption is that the spatial organization of refuges can be designed to ensure mating panmixia (that is, indiscriminate or random mating) and restriction of juvenile stages to just toxic or just non-toxic plants. Close intermixing of toxic and non-toxic plants may allow mobile individual insects to feed on a mixture, effectively diluting the toxin dose and compromising the effective recessivity of counter-resistance (Fig. 1b). By contrast, although large

© 2001 Macmillan Magazines Ltd

Box 2
Evolution of counter-resistance by haploid pests

A simple evolutionary model illustrates that a refuge strategy is not likely to delay the evolution of counter-resistance in haploid pests. This model assumes that there are two pest genotypes, one resistant to a genetically engineered plant toxin (*R*) and one susceptible to the toxin (*S*). There are also two crop varieties, one with the toxin and one without. The variety without the toxin is planted in refuges, which constitute a fraction β of the acreage planted. The relative fitness of the susceptible pest genotype is assumed to be 1 on plants without the toxin, and 0 on plants with the toxin. The relative fitness of the resistant pest genotype is taken to be $1-c$ on both crop varieties, where c represents the cost of counter-resistance. Such costs typically range from 0 to 0.2[66,68,104].

Using the general equation for change in allele frequencies in haploid populations[103], the equation that describes the change in frequency, p_S, of the resistant pest genotype from one generation to the next is

$$p'_R = (1-c)p_R / [(1-c)p_R + \beta]$$

This equation can be iterated, starting with an initially low frequency of the resistant genotype ($p_R = 0.0001$), to determine how the rate of spread of the resistant genotype depends on the size of the refuge and on the magnitudes of cost of resistance. Box 2 Table below shows the number of generations needed for the resistant genotype to become common (that is, to reach a frequency of 0.5).

These results indicate that, unless refuges constitute roughly half the acreage planted and costs of counter-resistance are very high, a refuge strategy is not likely to delay the evolution of counter-resistance substantially. Growers typically will not accept refuges that constitute more than 10% of the acreage planted, and the US Environmental Protection Agency mandates a refuge of 4% for transgenic resistant cotton[60]. With refuges of these sizes, substantial counter-resistance is likely to evolve within a few years in haploid pests.

Box 2 Table Number of generations to reach $p_R = 0.5$

β	c	With cost	Without cost
0.1	0.1	4	3
0.1	0.2	4	3
0.1	0.3	4	3
0.3	0.1	9	8
0.3	0.2	10	8
0.3	0.3	11	8
0.5	0.1	16	14
0.5	0.2	20	14
0.5	0.3	28	14

refuge and non-refuge patches may ensure that individuals will remain in the same patch throughout development, mating panmixia may be compromised if the patches are too large, leading to a preponderance of within-patch mating. Such assortative mating produces relatively more counter-resistant homozygotes and fewer heterozygotes, reducing the effectiveness of recessivity in slowing the evolution of counter-resistance (Fig.1c). Although some recent theoretical analyses have attempted to assess the effects of these complications on the effectiveness of the HDR strategy, detailed empirical investigations of insect movement patterns are necessary to determine whether, for any specific crop, an appropriate refuge configuration can be designed that will allow the HDR strategy to be effective[74].

Limitations of the high-dose/refuge strategy

The HDR strategy has been developed largely in the context of delaying the evolution of counter-resistance by sexually reproducing insects. But evolutionary considerations suggest that this strategy may be ineffective in managing counter-resistance in other types of organisms. Panmictic sexual reproduction is crucial for this strategy because it ensures that most copies of the initially rare counter-resistance allele occur in low-fitness heterozygotes. In diploid pests with substantial asexual reproduction (for example, aphids), a rare mutant homozygote can rapidly multiply on resistant hosts, increasing the frequency of counter-resistance.

For haploid pests (for example, viruses, bacteria and some fungi), achieving effective recessivity is by definition impossible. Consequently, slowing the spread of a counter-resistance allele within the HDR paradigm can be achieved only by providing a refuge, which reduces the fitness difference between counter-resistant and non-counter-resistant genotypes. Unfortunately, a simple evolutionary model suggests that even with a refuge constituting 50% of all plants, far greater than is commercially acceptable[60], and a large cost associated with counter-resistance, evolution of counter-resistance will not be substantially delayed (Box 2). It is therefore likely that alternate strategies will be necessary for resistance management in systems with haploid pests.

One augmentation of the HDR approach that has been examined theoretically is pyramiding — engineering plants to produce two unrelated toxins simultaneously. Models with pyramiding indicate that, compared to crops with just one toxin, much smaller refuges are required to achieve the same delay in pest adaptation[75–77]. As with the HDR approach with a single toxin, recessivity is critical in these models, because the added effectiveness of pyramiding is due largely to the rarity of doubly counter-resistant homozygote genotypes in the pest. With high doses of both toxins, such recessivity can theoretically be achieved, but the technical and economic difficulties of pyramiding even two unrelated toxins[78] make this strategy unfeasible for the foreseeable future.

One possible exception to this pessimistic conclusion is suggested by the recent molecular dissection of receptor genes for pathogen resistance (see above, and the review in this issue by Dangl and Jones, pages 826–833). The presence of multiple copies of these genes presumably increases the effectiveness of this type of defence in nature by increasing the likelihood that mutations will produce at least one receptor protein that can recognize a virulent pathogen. This multiplicity could also be used to advantage for resistance management by pyramiding five or six receptors that recognize different pathogen elicitors. Evolutionary stability of resistance would be conferred by the redundancy of the receptors — mutations would be required simultaneously in five or six pathogen elicitor molecules to confer counter-resistance (escape from recognition), an event with infinitesimally low probability. Moreover, by targeting as elicitors pathogen molecules that are involved in vital pathogen life processes, mutations in individual elicitors that render them no longer recognized by the corresponding receptor are likely to be detrimental. This cost would tend to prevent loss of elicitor recognition by genetic drift in the pathogen. In addition, such targeting would minimize the chance that some naturally occurring pathogen strains lack the elicitors, and thus would be virulent on the genetically modified crop.

Several considerations suggest that this type of pyramiding is likely to be more easily achieved than pyramiding unrelated toxins. First, only a single gene — that coding for the receptor — needs be inserted for each resistance factor, as different receptors all initiate the same signal cascade[79]. By contrast, in many cases, inserting new toxins will require inserting the genes coding for all of the enzymes required to make that toxin. (Bt-toxin is unusual in that it is the product of a single gene.) Second, the risk of increased autotoxicity associated with multiple toxins does not arise with receptor proteins, because they are not toxic to the plants that produce them. Finally, each toxin inserted into a plant potentially reduces yield because of costs associated with resistance. By contrast, if engineered receptors can be substituted for the redundant receptors already in a plant genome, it may be possible to completely avoid any costs of resistance.

© 2001 Macmillan Magazines Ltd

Of course, there still remain formidable obstacles to implementing this type of strategy. Among these are designing or discovering receptors that target a specific pathogen, and developing techniques for replacing native receptor genes with engineered genes. Nevertheless, the potential payoff of indefinitely lasting resistance to pathogens suggests that this is a strategy to pursue.

Tolerance as an evolutionarily stable defence

In developing approaches for managing resistance, most effort has focused on slowing artificial co-evolution by delaying the evolution of counter-resistance. However, recent investigations of plant defences in nature suggest that an alternate management strategy may be effective in some cases: breaking the co-evolutionary cycle by incorporating tolerance, rather than resistance, into crops.

Whereas resistance reduces the amount of damage or infection a plant experiences, tolerance reduces or eliminates the detrimental effect of a given amount of damage or infection on plant fitness (or on crop yield, in an agricultural context). Agricultural scientists recognized decades ago that crop cultivars could differ in tolerance, and have made some attempts to breed tolerance into crops[80,81]. But only over the past decade have evolutionary biologists discovered that natural plant populations often use tolerance as a defence against natural enemies. Genetic variation for degree of tolerance has been detected in plant families as taxonomically disparate as Piperaceae, Convolvulaceae, Polemoniaceae and Brassicaceae. Within species, populations often diverge in mean level of tolerance, and in some cases it is known that populations with a history of higher herbivore damage have higher tolerance[82]. In most plants, tolerance does not completely prevent damage from decreasing fitness, probably because costs prevent the evolution of maximum levels of tolerance[28]. Nevertheless, some species exhibit overcompensation, a form of tolerance in which enemy damage actually increases plant fitness[83,84].

In contrast to resistance, tolerance is not believed to adversely affect natural enemies[85–87]. Consequently, the evolution of tolerance does not generate natural selection for counter-adaptation in enemies, and thus breaks the co-evolutionary cycle. These considerations suggest that if crops could be genetically engineered to be tolerant to pests, counter-resistance management might no longer be an issue.

Many obstacles stand in the way of realizing this promise. The most important is that genes involved in conferring tolerance have not been identified at the molecular level for any plant species. A second is that conventional breeding programmes suggest that tolerance may often be a genetically complex trait involving many different plant characters[82]. A final obstacle is that increased tolerance may not prevent unacceptable cosmetic damage to a crop. Despite these obstacles, there is no reason to believe that some crops could not eventually be genetically engineered to be tolerant. As in the case of pyramiding receptor genes, the payoff of a possibly indefinitely stable defence should provide a strong economic incentive for funding the basic research needed to achieve this goal.

Coupling a toxin with non-preference

Approximately 90% of all herbivorous insects have narrow diets, feeding on plants of only one taxonomic family, and many species are confined to a single host species[22,88–90]. Both empirical and theoretical investigations suggest two general reasons for evolving such specialization: (1) variability in fitness on different host plant species favours behavioural genotypes that restrict feeding to the best hosts; and (2) maintaining mechanisms to nullify the disparate defensive adaptations of many different plant species is too costly for generalist herbivores[7].

By contrast, the key to understanding why specialized herbivores remain specialized lies in the observation that specialization involves both behavioural and physiological adaptation. Selection for behavioural genotypes that restrict feeding to a small number of plants increases selection for physiological adaptation to those plants and relaxes selection for physiological adaptation to other (non-host) plants[91,92]. Specialist species thus tend to have low fitness on non-host plants even if they can be induced to feed on them.

Population genetic models indicate that once a herbivore has evolved both behavioural and physiological specialization, the subsequent evolution of a broader diet is likely to be very difficult; if costs are associated with physiological adaptation to the novel host, specialization can be an evolutionarily stable state[91,93]. These theoretical considerations suggest another approach for resistance management: combining the HDR strategy with manipulation of the attractiveness of the toxic crop variety. Under this approach, a crop would be genetically engineered not only to produce a high dose of a toxin, but also to be unrecognized as a potential host by the pest. As in the HDR strategy, a refuge is provided that consists of a non-resistant variety of the crop or an alternative host.

Lack of recognition has long been recognized as a form of resistance, and has been bred into some crops[80], although by itself it is not evolutionarily stable. In many insects, behavioural recognition is based on one or a few 'token stimuli', often plant secondary compounds that stimulate insect feeding or oviposition[94]. In theory, genetically engineering non-preference resistance could in many cases involve simply deactivating one gene coding for an enzyme in the biochemical pathway that produces the token stimulus. Such a manipulation is likely to be simpler than inserting, under a pyramiding strategy, a new toxin that is biochemically even moderately complex to produce, and is likely to be equally effective.

Applied evolutionary research

Although the promise of genetically engineered crops that are resistant to important pests is beginning to be realized (see review in this issue by Stuiver and Custers, pages 865–868), the long-term success of this approach will be determined by economic realities. Because research and development costs are enormous compared to conventional breeding, their widespread use will depend on their having a reasonably long effective lifetime. And because the main threat to longevity is evolutionary change in the targeted pest species, management approaches that actively manipulate the evolutionary process will be required. Although evolutionary biologists have begun to develop approaches that allow such manipulation, most work so far has used simple evolutionary models with possibly unrealistic assumptions about the basic biology and genetics of the targeted pests. Moreover, there is currently little empirical evidence indicating whether the evolution of counter-resistance can actually be slowed or prevented as the theory suggests.

As the examples above illustrate, the continued study of evolution in natural plant–enemy systems is likely to contribute new insights regarding approaches to resistance management. At the same time, refining and improving current strategies such as HDR will require the development of more realistic and sophisticated population genetic models, the conceptual foundation of evolutionary biology. Yet the number of scientists engaged in these activities is small, compared with the number engaged in elucidating the molecular and biochemical causes of resistance. This dilemma should provide sufficient justification to stimulate the private sector, governments and universities to establish and fund a new initiative aimed at fostering increased research effort not only in the area of resistance management, but also in all applied disciplines that involve a significant evolutionary component.

The ultimate irony, of course, is that a call for enhanced funding for applied evolutionary research comes at a time of renewed anti-evolutionary religious zeal. In the United States, citizens in many of the very states whose economic welfare depends on crops susceptible to attack by devastating pests are calling for restricting or banning the teaching of the very science that holds out the most promise for winning the co-evolutionary war between crop plants and their enemies. We can only hope that continued education about the practical implications of evolutionary biology will persuade most citizens to ignore these calls. □

© 2001 Macmillan Magazines Ltd

1. Ehrlich, P. R. & Raven, P. H. Butterflies and plants: a study in coevolution. *Evolution* **18,** 586–608 (1964).
2. Thompson, J. N. *The Coevolutionary Process* (Univ. of Chicago Press, Chicago, 1994).
3. Whittaker, R. H. & Feeny, P. P. Allelochemics: chemical interactions between species. *Science* **171,** 757–770 (1971).
4. Janzen, D. H. When is it coevolution? *Evolution* **34,** 611–612 (1980).
5. Gould, F. in *Chemical Mediation of Coevolution* (ed. Spencer, K. C.) 13–55 (Academic Press, New York, 1988).
6. Berenbaum, M. R. & Zangerl, A. R. in *Chemical Mediation of Coevolution* (ed. Spencer, K. C.) 113–132 (Academic Press, New York, 1988).
7. Rausher, M. D. in *Evolutionary Perspectives in Insect Chemical Ecology* (eds Roitberg, B. D. & Isman, M. B.) 20–88 (Routledge, Chapman & Hall, New York, 1992).
8. Hatchett, J. H. & Gallun, R. L. Frequency of Hessian fly, *Mayteiola destructor*, races in field populations. *Ann. Entomol. Soc. Am.* **61,** 1446–1449 (1968).
9. Gallun, R. L. Genetic basis of Hessian fly epidemics. *Ann. NY Acad. Sci.* **287,** 223–229 (1977).
10. Foster, J. E., Ohm, H. W., Patterson, F. L. & Taylor, P. L. Effectiveness of deploying single gene resistances in wheat for controlling damage by the Hessian fly (Diptera: Cecidomyiidae). *Environ. Entomol.* **20,** 964–969 (1991).
11. Rosenthal, G. A., Dahlman, D. L. & Janzen, D. H. A novel means for dealing with L-canavanine, a toxic metabolite. *Science* **192,** 256–258 (1976).
12. Berenbaum, M. R. & Zangerl, A. R. Chemical phenotype matching between a plant and its insect herbivore. *Proc. Natl Acad. Sci. USA* **95,** 13743–13748 (1998).
13. Shirley, B. W. Flavonoid biosynthesis: 'new' functions for an old pathway. *Trends Plant Sci.* **11,** 377–382 (1996).
14. Muller, C. H. The "co-" in coevolution. *Science* **164,** 197–198 (1969).
15. Robinson, T. Metabolism and function of alkaloids in plants. *Science* **184,** 430–435 (1974).
16. Seigler, D. & Price, P. W. Secondary compounds in plants: primary functions. *Am. Nat.* **110,** 101–105 (1976).
17. Koes, R. E., Quattrocchio, F. & Mol, J. N. M. The flavonoid biosynthetic pathway in plants: function and evolution. *BioEssays* **16,** 123–132 (1994).
18. Seigler, D. S. Primary roles for secondary compounds. *Biochem. System. Ecol.* **5,** 195–199.
19. Jermy, T. Insect–host-plant relationship—co-evolution or sequential evolution? *Symp. Biol. Hung.* **16,** 109–113 (1976).
20. Jermy, T. Evolution of insect/host plant relationships. *Am. Nat.* **124,** 609–630 (1984).
21. Rausher, M. D. Genetic analysis of coevolution between plants and their natural enemies. *Trends Genet.* **12,** 212–217 (1996).
22. Strong, D. R., Lawton, J. H. & Southwood, R. *Insects on Plants* (Harvard Univ. Press, Cambridge, MA, 1984).
23. Bernays, E. & Graham, M. On the evolution of host specificity in phytophagous arthropods. *Ecology* **69,** 886–892 (1988).
24. Mauricio, R. & Rausher, M. D. Experimental manipulation of putative selective agents provides evidence for the role of natural enemies in the evolution of plant defense. *Evolution* **51,** 1435–1444 (1997).
25. Schonle, I. & Bergelson, J. Evolutionary ecology of the tropane alkaloids of *Datura stramonium* L. (Solanaceae). *Evolution* **54,** 778–788 (2000).
26. Biere, A. & Antonovics, J. Sex-specific costs of resistance to the fungal pathogen *Ustilago violacea* (*Microbotryum violaceum*) in *Silene alba*. *Evolution* **50,** 1098–1110 (1996).
27. Simms, E. L. & Rausher, M. D. The evolution of resistance to herbivory in *Ipomoea purpurea*. II. Natural selection by insects and costs of resistance. *Evolution* **43,** 573–585 (1989).
28. Tiffin, P. & Rausher, M. D. Genetic constraints and selection acting on tolerance to herbivory in the common morning glory, *Ipomoea purpurea*. *Am. Nat.* **154,** 700–716 (1999).
29. Parker, J. E. & Coleman, M. J. Molecular intimacy between proteins specifying plant-pathogen recognition. *Trends Biochem. Sci.* **22,** 291–296 (1997).
30. Hammond-Kosack, K. E. & Jones, J. D. G. Resistance gene-dependent plant defense responses. *Plant Cell* **8,** 1773–1791 (1996).
31. De Wit, P. J. G. M. Pathogen avirulence and plant resistance: a key role for recognition. *Trends Plant Sci.* **2,** 452–458 (1997).
32. Parniske, M. *et al.* Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the *Cf-4/9* locus of tomato. *Cell* **91,** 821–832 (1997).
33. Wang, G.-L. *et al. Xa21D* encodes a receptor-like molecular with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. *Plant Cell* **10,** 765–779 (1998).
34. Meyers, B. C., Shen, K. A., Rohani, P., Gaut, B. S. & Michelmore, R. W. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. *Plant Cell* **11,** 1833–1846 (1998).
35. McDowell, J. M. *et al.* Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis. Plant Cell* **10,** 1861–1874 (1998).
36. Li, W.-H. *Molecular Evolution* (Sinauer, Sunderland, MA, 1997).
37. Bishop, J. G., Dean, A. M. & Mitchell-Olds, T. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl Acad. Sci. USA* **10,** 5322–5327 (2000).
38. Simms, E. L. in *Plant Resistance to Herbivores and Pathogens: Ecology, Evolution and Genetics* (eds Fritz, R. S. & Simms, E. L.) 392–425 (Univ. of Chicago Press, Chicago, 1992).
39. Fagerstrõm, T., Larsson, S. & Tenow, O. On optimal defense in plants. *Funct. Ecol.* **1,** 73–81 (1987).
40. Simms, E. L. & Rausher, M. D. Costs and benefits of plant defense to herbivory. *Am. Nat.* **130,** 70–581 (1987).
41. Berenbaum, M. R., Zangerl, A. R. & Nitao, J. K. Constraints on chemical coevolution: wild parsnips and the parsnip webworm. *Evolution* **40,** 1215–1228 (1996).
42. Bergelson, J. The effect of genotype and the environment on costs of resistance in lettuce. *Am. Nat.* **143,** 349–359 (1994).
43. Fineblum, W. L. & Rausher, M. D. Tradeoff between resistance and tolerance to herbivore damage in a morning glory. *Nature* **377,** 517–520 (1995).
44. Mauricio, R. Costs of resistance to natural enemies in field populations of the annual plant, *Arabidopsis thaliana*. *Am. Nat.* **151,** 20–28 (1998).
45. Bergelson, J. & Purrington, C. B. Surveying patterns in the cost of resistance in plants. *Am. Nat.* **148,** 536–558 (1996).
46. Haughn, G. W. & Somerville, C. Sulfonylurea-resistant mutants of *Arabidopsis thaliana*. *Mol. Gen. Genet.* **204,** 430–434 (1986).
47. Haughn, G. W., Smith, J., Mazur, B. & Somerville, C. Transformation with a mutant *Arabidopsis* acetolactate synthase gene renders tobacco resistant to sulfonylurea herbicides. *Mol. Gen. Genet.* **211,** 266–271(1988).
48. Bergelson, J., Purrington, C. B., Palm, C. J. & Lopez-Guttierrez, J. C. Costs of resistance: a test using transgenic *Arabidopsis thaliana*. *Proc. R. Soc. Lond. B* **263,** 1659–1663 (1996).
49. Purrington, C.B. & Bergelson, J. Exploring the physiological basis of costs of herbicide resistance in *Arabidopsis thaliana*. *Am. Nat.* **154,** S82–S91 (1999).
50. Ewald, P. W. *Evolution of Infectious Disease* (Oxford Univ. Press, New York, 1994).
51. Levin, B. R., Lipsitch, M. & Bonhoeffer, S. Population biology, evolution, and infectious disease: convergence and synthesis. *Science* **283,** 806–809 (1999).
52. Law, R. & Grey, D. R. Evolution of yields from populations with age-specific cropping. *Evol. Ecol.* **3,** 343–359 (1989).
53. Heino, M. Management of evolving fish stocks. *Can. J. Fish. Aquat. Sci.* **55,** 1971–1982 (1998).
54. Murphy, E. J. & Rodhouse, P. G. Rapid selection effects in a short-lived semelparous squid species exposed to exploitation: inferences from the optimisation of life-history functions. *Evol. Ecol.* **13,** 517–537 (1999).
55. Hoffmann, A. A. & Parsons, P. A. *Evolutionary Genetics and Environmental Stress* (Oxford Univ. Press, New York, 1991).
56. Meefe, G. K. & Carroll, C. R. *Principles of Conservation Biology* 2nd edn (Sinauer, Sunderland, MA, 1997).
57. Landweber, L. F. & Pokrovskaya, I. D. Emergence of a dual-catalytic RNA with metal-specific cleavage and ligase activities: the spandrels of RNA evolution. *Proc. Natl Acad. Sci. USA* **96,** 173–178 (1999).
58. Landweber, L. F. Experimental RNA evolution. *Trends Ecol. Evol.* **14,** 353–358 (1999).
59. Proceedings of the Fourth International Meeting on DNA based Computers. *Biosystems* **52** (1999).
60. Gould, F. Sustainability of transgenic insecticidal cultivars: integrating pest genetics and ecology. *Annu. Rev. Entomol.* **43,** 701–726 (1998).
61. van Emden, H. F. Plant insect relationships and pest control. *World Rev. Pest Control* **5,** 115–123 (1966).
62. Roush, R. T. in *Advances in Insect Control: The Role of Transgenic Plants* (eds Carozzi, N. & Koziel, M.) 271–294 (Taylor & Francis, London, 1997).
63. Environmental Protection Agency. Plant pesticides resistance management. *Fed. Regist.* **62(36)** (1997).
64. Fischoff, D. A. Management of lepidopteran pests with insect resistant cotton: recommended approaches. *Proc. Beltwide Cotton Res. Conf. Natl Cotton Counc. Am., Memphis, TN*, 751–753 (1992).
65. Fischoff, D. A. in *Biotechnology and Integrated Pest Management* (ed. Persley, G. J.) 214–227 (CAB Int., Oxon, 1996).
66. Roush, R. T. & Plapp, F. W. Effects of insecticide resistance on biotic potential of the house fly (*Musca domestica*) (Diptera: Muscidae). *J. Econ. Entomol.* **75,** 708–713 (1982).
67. Fry, J. D. Trade-offs in fitness on different hosts: evidence from a selection experiment with a phytophagous mite. *Am. Nat.* **136,** 569–580 (1990).
68. Groeters, F. R., Tabashnik, B. E., Finson, N. & Johnson, M. W. Fitness costs of resistance to *Bacillus thuringiensis* in the diamondback moth (*Plutella xylostella*). *Evolution* **48,** 197–201 (1994).
69. Wood, R. J. & Bishop, J. A. in *Genetic Consequences of Man Made Change* (eds Bishop, J. A. & Cook, L. M.) 53–96 (Academic, London, 1981).
70. Hamma, H., Suzuki, K. & Yanaka, H. Inheritance and stability of resistance to *Bacillus thuringiensis* formulations of the diamondback moth, *Plutella xylostella* (Linnaeus) (Lepidoptera: Yponomeutridae). *Appl. Entomol. Zool.* **27,** 355–362 (1992).
71. Tabashnik, B. E., Finson, N., Groeters, F. R., Moar, W. J. & Johnson, M. W. Reversal of resistance to *Bacillus thuringiensis* in *Plutella xylostella*. *Proc. Natl Acad. Sci. USA* **91,** 4120–4124 (1994).
72. Perez, C. J., Shelton, A. M. & Roush, R. T. Managing diamondback moth (Lepidoptera: Plutellidae) resistance to folar applications of *Bacillus thuringiensis*: testing strategies in field cages. *J. Econ. Entomol.* **90,** 1462–1470 (1997).
73. Liu, Y. B. & Tabashnik, B. E. Experimental evidence that refuges delay insect adaptation to *Bacillus thuringiensis*. *Proc. R. Soc. Lond. B* **264,** 605–610 (1997).
74. Shelton, A. M., Tang, J. D., Roush, R. T., Metz, T. D. & Earle, E. D. Field tests on managing resistance to Bt-engineered plants. *Nature Biotechnol.* **18,** 339–342 (2000).
75. Cox, T. S. & Hatchett, J. H.. Genetic model for wheat/Hessian fly (Diptera: Cecidomyiidae) interaction: strategies for deployment of resistance genes in wheat cultivars. *Environ. Entomol.* **15,** 24–31 (1986).
76. Gould, F. The evolutionary potential of crop pests. *Am. Sci.* **79,** 496–507 (1991).
77. Roush, R. T. Two-toxin strategies for management of insecticidal transgenic crops: can pyramiding succeed where pesticide mixtures have not? *Phil. Trans. R. Soc. Lond. B* **353,** 1777–1786 (1998).
78. Roush, R. T. Bt-transgenic crops: just another pretty insecticide or a chance for a new start in resistance management? *Pesticide Sci.* **51,** 328–334 (1997).
79. Pryor, T. The origin and structure of fungal disease resistance in plants. *Trends Genet.* **3,** 157–161 (1987).
80. Painter, R. *Insect Resistance in Crop Plants* (Wiley, New York, 1951).
81. Schafer, J. Tolerance to plant disease. *Annu. Rev. Phytopathol.* **9,** 235–252 (1971).
82. Stowe, K. A., Marquis, R. J., Hochwender, C. G. & Simms, E. L. The evolutionary ecology of tolerance to consumer damage. *Annu. Rev. Ecol. Syst.* **31,** 565–595 (2000).
83. Paige, K. N. & Whitham, D. G. Overcompensation in response to mammalian herbivory: the advantage of being eaten. *Am. Nat.* **143,** 739–749 (1987).
84. Agrawal, A. A. Overcompensation of plants in response to herbivory and the by-product benefits of mutualism *Trends Plant Sci.* **5,** 309–313 (2000).
85. Rosenthal, J. P. & Kotanen, P. M. Terrestrial plant tolerance to herbivory. *Trends Ecol. Evol.* **9,** 145–148 (1994).
86. Strauss, S. & Agrawal, A. The ecology and evolution of tolerance to herbivory. *Trends Ecol. Evol.* **14,** 179–185 (1999).
87. Tiffin, P. Are tolerance, avoidance and antibiosis evolutionarily and ecologically equivalent responses of plants to herbivores? *Am. Nat.* **155,** 128–138 (2000).
88. Futuyma, D. J. & Gould, F. Associations of plants and insects in a deciduous forest. *Ecol. Monogr.* **49,** 33–50 (1979).
89. Thompson, J. N. *Interaction and Coevolution*. (Wiley, New York, 1982).
90. Bernays, E. & Graham, M. On the evolution of host specificity in phytophagous arthropods. *Ecology* **69,** 886–892 (1988).
91. Rausher, M. D. in *Evolution of Insect Pests: The Pattern of Variations* (eds Kim, K. C & McPheron, B. A) 259–283 (Wiley, New York, 1993).
92. Fry, J. D. The evolution of host specialization: are trade-offs overrated? *Am. Nat.* **148,** S84–S107 (1996).

© 2001 Macmillan Magazines Ltd

93. Castillo-Chavez, C., Levin, S. A. & Gould, F. Physiological and behavioral adaptation to varying environments: a mathematical model. *Evolution* **42,** 986–994 (1988).
94. Feeny, P. in *Herbivores: Their Interactions with Secondary Plant Metabolites* (eds Rosenthall, G. A. & Berenbaum, M.) 1–44 (Academic, San Diego, 1992).
95. Berenbaum, M. Coumarins and caterpillars: a case for coevolution. *Evolution* **37,** 163–179 (1983).
96. Cohen, M. B., Schuler, M. A. & Berenbaum, M. R. A host-inducible cytochrome P-450 from a host-specific caterpillar: molecular cloning and evolution. *Proc. Natl Acad. Sci. USA* **89,** 10920–10924 (1992).
97. Rathcke, B. J. & Poole, R. W. Coevolutionary race continues: butterfly larval adaptation to plant trichomes. *Science* **187,** 175–176 (1975).
98. Dussourd, D. E. & Eisner, T. Vein-cutting behavior: insect counterploy to the latex defense of plants. *Science* **237,** 898–901 (1987).
99. Carroll, S. P. & Loye, J. E. Specialization of *Jadera* species (Himiptera: Rhopalidae) on the seeds of Sapindaceae (Sapindales), and coevolutionary responses of defense and attack. *Ann. Entomol. Soc. Am.* **80,** 373–387 (1987).
100. Carroll, S. P. & C. Boyd. Host race radiation in the soapberry bug: natural history with the history. *Evolution* **46,** 1052–1069 (1992).
101. Janzen, D. H. Coevolution of mutualism between ants and acacias in Central America. *Evolution* **20,** 249–275 (1966).
102. Eubanks, M. D., Nesci, K. A., Petersen, M. K., Liu, Z. & Sanchez, H. B. The exploitation of an ant-defended host plant by a shelter-building herbivore. *Oecologia* **109,** 454–460 (1997).
103. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 2nd edn (Sinauer, Sunderland, MA, 1989).
104. Eggers-Schumacher, H. A. A comparison of the reproductive performance of insecticide-resistant and susceptible clones of *Myzus persicae*. *Entomol. Exp. Appl.* **34,** 301–307 (1983).

© 2001 Macmillan Magazines Ltd

Engineering disease resistance in plants

Maarten H. Stuiver & Jerome H. H. V. Custers

Syngenta-MOGEN, Einsteinweg 97, 2333 CB Leiden, The Netherlands (e-mail: Maarten.Stuiver@syngenta.com)

Ever since the initial discovery of the molecules and genes involved in disease resistance in plants, attempts have been made to engineer durable disease resistance in economically important crop plants. Unfortunately, many of these attempts have failed, owing to the complexity of disease-resistance signalling and the sheer diversity of infection mechanisms that different pathogens use. Although disease-resistant transgenic plants or seeds are not yet available commercially, future product development seems likely as our current level of understanding of pathogenesis and plant defence improves.

One cannot study plant defence without being impressed by the complex and sophisticated systems that plants have evolved to withstand a variety of pathogens. It is often mentioned that plants successfully withstand infection by the vast majority of pathogens that attack them. It is the sheer diversity of the infection mechanisms that these pathogens use that makes this feat truly remarkable.

Plant defence must be adapted to combat two different types of pathogen. Necrotrophs are pathogens that produce toxic enzymes and metabolites that kill the tissue directly upon invasion. In contrast, hemibiotrophs or biotrophs initially feed on plants parasitically, keeping the cells in infected plant tissue alive for a significant fraction of the pathogen's life cycle; this is sometimes followed by a more necrotrophic existence during the later stages of infection. The number of different plant toxic compounds and proteins that have been isolated from plant pathogens and that contribute to virulence is enormous[1,2].

Using defence pathways to engineer resistance

During the past few years, the identification of key regulatory genes in plant defence has provided evidence that plants use several different defence pathways against different pathogens[3–5]. In general, these pathways are characterized by the signalling molecules that are crucial in the regulation of expression of defence proteins. The best-known signalling molecule is salicylic acid. Treatment of plants with salicylic acid or analogues of salicylic acid induces expression of a subset of plant defence responses[6], which results in warding off certain, but not all, pathogens[7,8]. Other pathways seem to use the small signalling molecules jasmonic acid and ethylene for their activation. Activation of either of these pathways also induces resistance, but to a different group of pathogens than that associated with salicylic acid[8,9]. In addition to these three pathways, evidence suggests that additional signalling molecules might be involved, most likely including reactive oxygen species[10,11].

Treatment of plants with one or more of these signalling molecules causes the coordinated induction of antifungal proteins, phytoalexins, enzymes involved in plant cell-wall reinforcement or breakdown of pathogen infection structures. Most known antifungal proteins identified from pathogen-infected plants[12] can be induced by either of these stimuli, alone or in combination[13], indicating that at least the crude outline of the plant defence potential has been established.

The wide spectrum of defence responses caused by treatment with each of these signalling molecules has prompted research to identify and use signal-transduction 'master switches' to engineer disease resistance. This has been successful in some cases. The *Arabidopsis NIM1/NPR1* gene seems to be crucial in salicylic acid-mediated resistance, and overexpression leads to resistance against several pathogens[14]. Other enhanced disease resistance (*edr*) mutants have been identified[15,16]. But engineering resistance through use of these master switches is generally not without drawbacks. Most mutants possessing constitutive expression of a defence pathway show reduced yield or plant vigour. And there seems to be antagonism between the different defence pathways[17,18], which leads to increased susceptibility to other pathogens[19].

But these effects may not prevent this approach from being taken to engineer resistance. Plants have a finely tuned defence (both with respect to resistance to different classes of pathogens and to the amount of energy devoted to pathogen resistance and yield), which has contributed to survival in the wild. In modern agriculture, it may be acceptable to shift this balance to provide resistance to the most problematic pathogens, with the threat posed by other pathogens being removed by adoption of, for example, growing conditions. Barley culture provides a good example of this technique. Most modern barley cultivars possess the naturally occurring mutant *mlo* gene, which provides durable and broad-spectrum resistance to powdery mildew, one of the main pathogens in barley. The use of the *mlo* gene leads to a small decrease in yield (P. Schulze-Lefert, personal communication), and makes barley more susceptible to another fungal pathogen, *Magnaporthe grisea*[20]. But this pest is not a significant problem in modern barley culture. The loss in yield is also acceptable, as yield stability (the ability to harvest a stable yield from year to year) is much more important to growers than absolute yield, especially for crops in which diseases can cause heavy losses. The possibility of using key regulators of defence pathways to 'tweak' resistance to the most pressing problems in agriculture provides an exciting opportunity, which is now starting to be realized.

One of the oldest strategies in the engineering of pathogen resistance is the overexpression of antifungal (or antipathogenic) proteins. In some respects this is very

© 2001 Macmillan Magazines Ltd

similar to the pathway-modulating approach described above, only much more specific, as only one or a few genes from the entire defence system are transferred simultaneously to a new transgenic crop. In contrast to the pathway-modulating approach, the impact on yield or the interference/antagonism with other defence pathways is most likely to be limited or absent. The limitation of this approach is that in many cases it will be highly specific for only a few pathogens, and generally it does not provide broad-spectrum control[21]. These constraints, however, do not undermine the usefulness of this strategy, as often only a small number of pathogens are truly important per crop.

One of the more significant practical problems encountered with this strategy is that the effect of the antifungal proteins is influenced by the endogenous defence mechanism already present in wild-type plants — the newly introduced proteins have to fit with the plant's endogenous defence compounds. And although all plants studied so far seem to have defence systems induced by salicylic acid, jasmonic acid and ethylene, the effector antipathogenic proteins and compounds differ considerably. In an elegant demonstration of the difficulty of this strategy, Punja and Raharjo[22] showed that transfer of a chitinase gene, which encodes a protein that degrades the cell wall of many fungi, to two different crops resulted in a resistance-elevating effect in carrot, but not in cucumber, even when the same pathogen was used to challenge the crops. But with the ongoing genome sequencing of key agricultural crops, and advances in the study of antimicrobial gene expression[23], it will probably be possible to identify the 'gaps' in pathogen defence of certain plant species, and so complement these systems more effectively.

Phytoalexins can be important in plant defence (see review in this issue by Dixon, pages 843–847), but in general the specific activity of these compounds is relatively limited, and where they are found to be a key factor in disease resistance (for example, resveratrol in grapes), the amounts accumulated are extremely high. Accumulation of such high levels of resveratrol is usually not possible when the appropriate genes required for synthesis are transferred to other crops[24]. Attempts to engineer resistance using this strategy have worked in tobacco and alfalfa[24–26], but the number of successes has remained low, and the level of resistance relatively modest. It is questionable whether this strategy will ever be widely used.

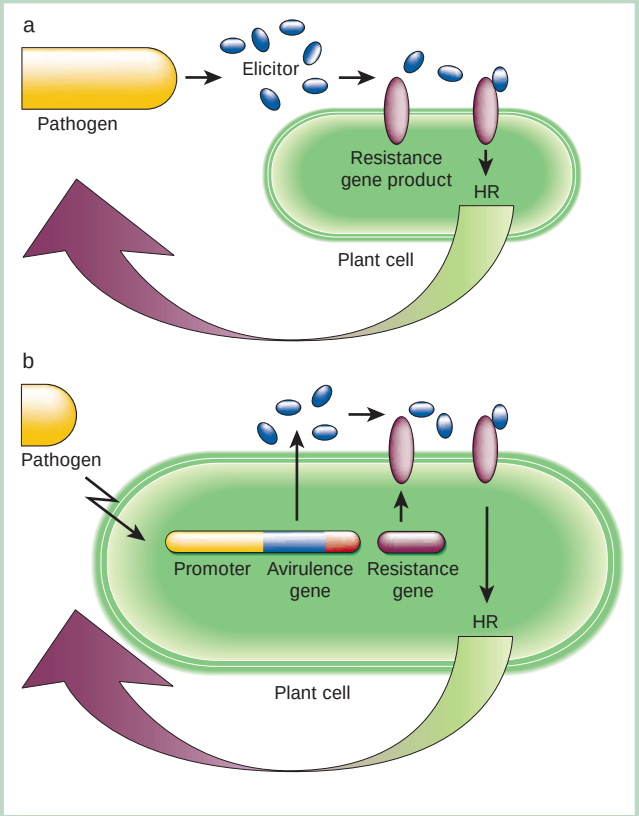


Figure 1 Generating broad-spectrum disease resistance using an elicitor and resistance gene. **a**, The hypersensitive response (HR) is triggered by the highly specific recognition of a pathogen-derived elicitor by a plant resistance gene product. The powerful and concerted defence (see text) that constitutes the hypersensitive response stops the pathogen. **b**, The components involved in the basic switch of the hypersensitive response can be used to create a more nonspecific defence system. A plant-derived pathogen-inducible promoter drives expression of a pathogen elicitor gene. The elicitor formed will trigger the hypersensitive response if the plants also contain the resistance gene.

Resistance genes and the hypersensitive response

The hypersensitive response is the most powerful defence system that plants have. It is a highly concerted (both temporally and spatially), complex defence response that involves local cell death, high local accumulation of phenolic compounds and cell-wall reinforcements in cells surrounding the area of cell death, and a more distal induction of general induced defence, which prevents further infections on distal parts of the plant (refs 27–30, and see review in this issue by Lam, Kato and Lawton, pages 848–853).

Although the defence response is powerful, and may stop infection by viruses, nematodes, bacteria and fungi, its limitations are that it is normally triggered only by the highly specific recognition (through means of a resistance gene) of a (pathogen-associated) elicitor molecule (see review in this issue by Dangl and Jones, pages 826–833). No response is induced even when closely related pathogens lacking this particular elicitor infect the plant.

Breeders have often used resistance genes to introduce resistance in their crops, and with a few exceptions, all introgressed resistance genes have been shown to lack durability in the field[31]. Pathogens are usually able to overcome resistance gene-mediated recognition either by shedding the corresponding elicitor gene, or by accumulating mutations in the gene, which prevents the gene product from being recognized, and thus fails to trigger the hypersensitive response[32,33].

The use of most known resistance genes is therefore limited in plant biotechnology for disease resistance. For those rare genes that can provide durable resistance, the possibility of transfer to commercially relevant crops is often limited, as resistance genes frequently fail to work when transferred between plant species, especially when the species are not closely related[34].

Non-host resistance

An exception to these pessimistic assessments occurs where resistance genes recognize molecules that are so important to the pathogen that they can neither be shed nor mutated. This may explain the ability of certain plant species to withstand infection to all known isolates of a given pathogen. There are doubtless several reasons why pathogens cannot infect most plant species, but in some cases of non-host resistance it has been shown that such a generic hypersensitive response provides the most important barrier to infection[35]. These findings have rekindled interest in the use of resistance genes to engineer pathogen resistance.

One of the most striking examples of non-host resistance based on the hypersensitive response is found in tobacco (*Nicotiana*) species against *Phytophthora infestans*, the causal agent of the devastating late blight in potato and tomato. The *Phytophthora*-derived INF1 gene product triggers a hypersensitive response reaction in these tobacco non-host plants, and targeted *INF1* deletions in *P. infestans* make the pathogen capable of infecting one of the tobacco species[36]. Although *INF1* deletion did not affect pathogenicity of *P. infestans* on potato, the deletion is found at an extremely low frequency, suggesting the gene has an important role for this pathogen. This in part might explain the durability of non-host resistance.

© 2001 Macmillan Magazines Ltd

The *Arabidopsis Eds1* gene is an essential component in the downstream signalling pathway mediated by resistance-gene loci conferring race-specific disease resistance to several *Peronospora parasitica* isolates and *Pseudomonas syringae* bacteria[37]. Mutation of the *Eds1* gene in *Arabidopsis thaliana* made the plants susceptible to the pathogen *Albugo candida* and several isolates of *P. parasitica*, which are not normally pathogenic on *Arabidopsis*, but infect *Brassica oleracea* subspecies (cabbages)[38]. This suggests that on non-mutated *Arabidopsis*, these pathogens can be prevented from infecting the plant through an effective hypersensitive response. If so, it should be possible to isolate such resistance genes and transfer them to other crops. But the question remains as to why such resistance gene-mediated resistance is more durable than that found normally in wild relatives and varieties used by breeders.

Generating broad-spectrum disease resistance

It is extremely tempting to make use of the hypersensitive response pathway and try to trigger this sort of defence after infection by different pathogens. Based on an idea originally advanced by Pierre de Wit[39], both our research team at Syngenta-MOGEN (in collaboration with de Wit's group; unpublished results) and the research group of Ricci and co-workers[40] have successfully engineered broad-spectrum disease resistance in plants. This involves transfer of a pathogen-derived elicitor gene to the plant, expression of which is made conditional on pathogen infection by putting it under control of a tightly regulated pathogen-inducible plant promoter (Fig. 1). Both teams have used this technology to create transgenic plants (tomato and tobacco, respectively) that show broad-spectrum and high-level fungal control, and our team has shown that the transgenic tomato plants can stop virus (tomato spotted wilt virus) infection.

The key issue in this strategy is the tight regulation of the pathogen-inducible promoter. The elicitor triggers local cell death and a large array of defence responses, and leakiness of the promoter can and will influence plant vigour and yield. Nevertheless, both teams have produced transgenic plants that show no sign of spontaneous firing of the hypersensitive response in the absence of pathogen challenge. This is therefore one of the more promising approaches to engineer broad-spectrum disease resistance using the endogenous defence components of plants.

Despite the huge induction of defence components that results from triggering the hypersensitive response, there are reports indicating that not all pathogens are stopped. Infection by *Botrytis cinerea*, a necrotrophic fungal pathogen of many plants, seems to be enhanced when a hypersensitive response is triggered in *Arabidopsis*[41] and tomato (Hennin *et al.*, unpublished results). We have also found no evidence of increased resistance (but also no evidence of increased susceptibility) to *Alternaria solani*, another necrotrophic fungal pathogen of tomato, although in the transgenic tobacco plants mentioned above[39], there is clearly increased resistance to this class of pathogens. Research by Bonnet *et al.*[42] indicates clearly that necrotrophic pathogens can be stopped by a hypersensitive response, so it seems that this strategy may have the potential to engineer resistance to necrotrophic pathogens.

Interference with pathogenesis

Most fungal and bacterial pathogens possess a diverse range of enzymes, proteins and metabolites that assist in the infection process of the plant. Although many such compounds seem to be dispensable for the pathogen without compromising its pathogenicity[43], key factors have been identified for a small number of pathogens that are crucial in the infection process[44–47]. The identification of these factors enables the design of strategies to neutralize them, and so interfere with pathogenesis.

For example, the necrotrophic fungus *Sclerotinia sclerotiorum* synthesizes large amounts of oxalic acid when infecting a plant. This compound is thought to be crucial in *Sclerotinia* infection, as mutants that loose the ability to synthesize oxalic acid are invariably non-pathogenic, whereas those that regain the ability to make this compound recover their virulence[48]. Oxalic acid serves as a co-factor needed to degrade plant cell-wall structures, which enables the fungus to colonize the plant, but it was also found recently to inhibit the onset of plant defence[48].

The dependence of the fungus on oxalic acid can be used to engineer efficient resistance[49]. Using oxalic acid oxidase, an enzyme naturally occurring in, for example, germinating wheat and barley (both plants have a significant level of resistance against *S. sclerotiorum*), the oxalate can be broken down to carbon dioxide and hydrogen peroxide, which disable the main fungal pathogenicity factor. The hydrogen peroxide may serve a second function, as it is a signalling molecule involved in inducing plant defence. Thus, oxalic acid oxidase serves a dual role: it disables fungal pathogenicity, and with the breakdown product it could boost plant defence[50]. Overexpression of another enzyme that breaks down oxalate, oxalate decarboxylase, also leads to increased resistance to *Sclerotinia* infection[51], although the reaction does not lead to accumulation of hydrogen peroxide, or other known signalling molecules. This suggests that the breakdown of the pathogenicity factor is sufficient to engineer resistance.

Proteins with a key role in pathogenicity have been discovered for other pathogens[33,52], and this is prompting strategies to counteract or interfere with their function[52]. Another interesting strategy that has recently been reported is interference with pathogen-induced apoptosis (programmed cell death). Several pathogens are known to induce apoptotic cell death through secreted toxin molecules, and interference with apoptosis through expression of apoptosis-inhibiting proteins makes these transgenic plants more resistant to several unrelated necrotrophic pathogens and isolated toxins[53,54]. It should be added, however, that current knowledge of the molecules involved in pathogenicity represents only the tip of the iceberg. For many fungi, even those of significant economic importance, the molecular basis of pathogenicity is poorly understood.

Looking forward

Engineering resistance to diseases has proven much more recalcitrant than to insects. Whereas a significant proportion of crops grown today has been engineered to express insect resistance (although exclusively through the use of *Bacillus thuringiensis* toxin genes), no commercial transgenic product with enhanced disease resistance is currently available. At this moment, our knowledge of pathogenesis and defence is still relatively primitive, but as engineering resistance becomes more sophisticated and successful in the coming years, we will no doubt see the first disease-resistant genetically modified crops appear on the market.

The current sequencing of crop plant genomes, together with comprehensive gene expression and functional gene analysis, will no doubt boost the development of transgenic disease-resistant plants. In particular, strategies using general defence pathways and antifungal protein overexpression will benefit from a thorough understanding of the defence arsenal of crop plants. It will provide indications of how to either complement defence or 'tweak' it. In addition, pathogen-responsive promoters, needed for development of a non-pathogen-specific hypersensitive response, will be readily identified. It is even possible that some non-host resistance genes might be uncovered, which can then be tested for durability.

Slightly lagging behind these developments is pathogen genome analysis. For several pathogens, sequencing of genomes or expressed sequence tags has already been initiated. Once pathogenicity genes have been identified, strategies aimed at interference with the disease process will become apparent. However, functional gene analysis (such as making knockout mutants) is technically challenging for some pathogens, such that this process might be relatively slow.

With increased public concerns about fungicide use on food crops, there is a clear need for alternatives, including biotechnology.

© 2001 Macmillan Magazines Ltd

Many of the described approaches can provide a high level of protection, and because they are based on a plant's own defence arsenal, they are likely to provide durable resistance as well. □

1. Walton, J. D. Host-selective toxins: agents of compatibility. *Plant Cell* **8,** 1723–1733 (1996).
2. White, F. F, Yang, B. & Johnson, L. B. Prospects for understanding avirulence gene function. *Curr. Opin. Plant Biol.* **3,** 291–298 (2000).
3. Delaney, T. P. *et al.* A central role of salicylic acid in plant disease resistance. *Science* **266,** 1247–1250 (1994).
4. Thomma, B. P., Penninckx, I. A., Broekaert, W. F. & Cammue, B. P. The complexity of disease signaling in Arabidopsis. *Curr. Opin. Immunol.* **13,** 63–68 (2001).
5. Pieterse, C. M. & van Loon, L. C. Salicylic acid-independent plant defence pathways. *Trends Plant Sci.* **4,** 52–58 (1999).
6. Uknes, S. *et al.* Acquired resistance in Arabidopsis. *Plant Cell* **4,** 645–656 (1992).
7. Gorlach, J. *et al.* Benzothiadiazole, a novel class of inducers of systemic acquired resistance, activates gene expression and disease resistance in wheat. *Plant Cell* **8,** 629–643 (1996).
8. Thomma, B. P. *et al.* Separate jasmonate-dependent and salicylate-dependent pathways in Arabidopsis are essential for resistance to distinct microbial pathogens. *Proc. Natl Acad. Sci. USA* **95,** 15107–15111 (1998).
9. Feys, B. J. & Parker, J. E. Interplay of signalling pathways in plant disease resistance. *Trends Genet.* **16,** 449–455 (2000).
10. Grant, J. J., Yun, B. W. & Loake, G. J. Oxidative burst and cognate redox signalling reported by luciferase imaging: identification of a signal network that functions independently of ethylene, SA and Me-JA but is dependent on MAPKK activity. *Plant J.* **24,** 569–582 (2000).
11. Pieterse, C. M. *et al.* A novel signaling pathway controlling induced systemic resistance in Arabidopsis. *Plant Cell* **10,** 1571–1580 (1998).
12. Stintzi, A. *et al.* Plant 'pathogenesis-related' proteins and their role in defense against pathogens. *Biochimie* **75,** 687–706 (1993).
13. Penninckx, I. A. M. A, Thomma, B. P. H. J., Buchala, A., Metraux, J.-P. & Broekaert, W. F. Concomitant activation of jasmonate and ethylene response pathways is required for induction of a plant defensin gene in Arabidopsis. *Plant Cell* **10,** 2103–2113 (1998).
14. Cao, H., Li, X. & Dong, X. Generation of broad-spectrum disease resistance by overexpression of an essential regulatory gene in systemic acquired resistance. *Proc. Natl Acad. Sci. USA* **95,** 6531–6536 (1998).
15. Bowling, S. A. *et al.* A mutation in Arabidopsis that leads to constitutive expression of systemic acquired resistance. *Plant Cell* **6,** 1845–1857 (1994).
16. Frye, C. A., Tang, D. & Innes, R. W. Negative regulation of defense responses in plants by a conserved MAPKK kinase. *Proc. Natl Acad. Sci. USA* **98,** 373–378 (2001).
17. Doares, S. H., Narvaez-Vazquez, J., Conconi, A. & Ryan, C. A. Salicylic acid inhibits synthesis of proteinase inhibitors in tomato leaves induced by systemin and jasmonic acid. *Plant Physiol.* **108,** 1741–1746 (1995).
18. Leslie, C. A. & Romani, R. J. Inhibition of ethylene biosynthesis by salicylic acid. *Plant Physiol.* **88,** 833–837 (1988).
19. Hoffman, T., Schmidt, J. S., Zheng, X. & Bent, A. F. Isolation of ethylene-insensitive soybean mutants that are altered in pathogen susceptibility and gene-for-gene disease resistance. *Plant Physiol.* **119,** 935–949 (1999).
20. Jarosch, B., Kogel, K.-H. & Schaffrath, U. The ambivalence of the barley *Mlo* locus: mutations conferring resistance against powdery mildew (*Blumeria graminis* f. sp. *hordei*) enhance susceptibility to the rice blast fungus *Magnaporthe grisea*. *Mol. Plant-Microbe Interact.* **12,** 508–514 (1999).
21. Alexander, D. *et al.* Increased tolerance to two oomycete pathogens in transgenic tobacco expressing pathogenesis-related proteon 1a. *Proc. Natl Acad. Sci. USA* **90,** 7327–7331 (1993).
22. Punja, Z. K. & Raharjo, S. H. T. Response of transgenic cucumber and carrot plants expressing different chitinase enzymes to inoculation with fungal pathogens. *Plant Dis.* **80,** 999–1005 (1996).
23. Maleck, K. *et al.* The transcriptome of *Arabidopsis thaliana* during systemic acquired resistance. *Nature Genet.* **26,** 403–410 (2000).
24. Hain, R. *et al.* Disease resistance results from foreign phytoalexin expression in a novel plant. *Nature* **361,** 153–156 (1993).
25. He, X. Z. & Dixon, R. A. Genetic manipulation of isoflavone 7-O-methyltransferase enhances biosynthesis of 4′-O-methylated isoflavonoid phytoalexins and disease resistance in alfalfa. *Plant Cell* **12,** 1689–1702 (2000).
26. Hipskind, J. D. & Paiva, N. L. Constitutive accumulation of a resveratrol-glucoside in transgenic alfalfa increases resistance to *Phoma medicaginis*. *Mol. Plant-Microbe Interact.* **13,** 551–562 (2000).
27. Dorey, S. *et al.* Spatial and temporal induction of cell death, defense genes, and accumulation of salicylic acid in tobacco leaves reacting hypersensitively to a fungal glycoprotein elicitor. *Mol. Plant-Microbe Interact.* **10,** 646–655 (1997).
28. Costet, L. *et al.* Relationship between localized acquired resistance (LAR) and the hypersensitive response (HR): HR is necessary for LAR to occur and salicylic acid is not sufficient to trigger LAR. *Mol. Plant-Microbe Interact.* **12,** 655–662 (1999).
29. Morel, J.-B. & Dangl, J. L. The hypersensitive response and the induction of cell death in plants. *Cell Death Differ.* **4,** 671–683 (1997).
30. Melchers, L. S. & Stuiver, M. H. Novel genes for disease-resistance breeding. *Curr. Opin. Plant Biol.* **3,** 147–152 (2000).
31. Pink, D. & Puddephat, I. Deployment of disease resistance genes by plant transformation—a 'mix and match' approach. *Trends. Plant Sci.* **4,** 71–75 (1999).
32. Joosten, M. H. A. J., Cozijnsen, T. J. & De Wit, P. J. G. M. Host resistance to fungal tomato pathogen lost by a single base pair change in an avirulence gene. *Nature* **367,** 384–386 (1994).
33. Gassmann, W. *et al.* Molecular evolution of virulence in natural field strains of *Xanthomonas campestris* pv. *vesicatoria*. *J. Bacteriol.* **182,** 7053–7059 (2000).
34. Rommens, C. M. & Kishore, G. M. Exploiting the full potential of disease-resistance genes for agricultural use. *Curr. Opin. Biotechnol.* **11,** 120–125 (2000).
35. Heath, M. C. Nonhost resistance and non-specific plant defenses. *Curr. Opin. Plant Biol.* **3,** 315–319 (2000).
36. Kamoun, S., van West, P., Vleeshouwers, V. G. A. A., de Groot, K. E. & Govers, F. Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of the elicitor protein INF1. *Plant Cell* **10,** 1413–1425 (1998).
37. Aarts, N. *et al.* Different requirements for EDS1 and NDR1 by disease resistance genes define at least two R gene-mediated signaling pathways in Arabidopsis. *Proc. Natl Acad. Sci. USA* **95,** 10306–10311 (1998).
38. Parker, J. E. *et al.* Characterization of eds1, a mutation in Arabidopsis suppressing resistance to Peronospora parasitica specified by several different RPP genes. *Plant Cell* **8,** 2033–2046 (1996).
39. de Wit, P. J. G. M. Molecular characterization of gene-for-gene systems in plant-fungus interactions and the application of avirulence genes in control of plant pathogens. *Annu. Rev. Phytopathol.* **30,** 391–418 (1992).
40. Keller, H. *et al.* Pathogen induced elicitin production in transgenic tobacco generates a hypersensitive response and nonspecific disease resistance. *Plant Cell* **11,** 223–235 (1999).
41. Govrin, E. M. & Levine, A. The hypersensitive response facilitates plant infection by the necrotrphic pathogen *Botrytis cinerea*. *Curr. Biol.* **10,** 751–757 (2000).
42. Bonnet, P., Bourdon, E., Ponchet, M., Blein, J. P. & Ricci, P. Acquired resistance triggered by elicitins in tobacco and other plants. *Eur. J. Plant Pathol.* **102,** 181–192 (1996).
43. Schafer, W. The role of cutinase in fungal pathogenesis. *Trends Microbiol.* **1,** 69–71 (1993).
44. Proctor, R. H., Hohn, T. M. & McCormick, S. P. Reduced virulence of *Gibberella zeae* caused by disruption of a trichothecene toxin biosynthetic gene. *Mol. Plant-Microbe Interact.* **8,** 593–601 (1995).
45. Johal, G. S. & Briggs, S. P. Reductase activity encoded by the Hm1 resistance gene in maize. *Science* **258,** 985–987 (1992).
46. Tanaka, A., Shiotani, H., Yamamoto, M. & Tsuge, T. Insertional mutagenesis and cloning of the genes required for biosynthesis of the host-specific AK-toxin in the Japanese pear pathotype of *Alternaria alternata*. *Mol. Plant-Microbe Interact.* **12,** 691–702 (1999).
47. Daub, M. E. & Ehrenshaft, M. The photoactivated *Cercospora* toxin cercosporin: contributions to plant disease and fundamental biology. *Annu. Rev. Phytopath.* **38,** 461–490 (2000).
48. Cessna, S. G., Sear, V. E., Dickman, M. B. & Low, P. S. Oxalic acid, a pathogenicity factor for *Sclerotinia sclerotiorum*, suppresses the oxidative burst of the host plant. *Plant Cell* **12,** 2191–2200 (2000).
49. Hartman, C. L., Sarjit, J. & Schmitt, M. S. Newly characterised oxalate oxidase and uses therefor. Patent application WO9214824 <http://ep.espacenet.com> (1992).
50. Charne, D. G. *et al.* Production of pathogen resistant plants. Patent application WO9904012 <http://ep.espacenet.com> (1999).
51. Kesarwani, M., Azam, M., Natarajan, K. & Datta, A. Oxalate decarboxylase from *Collybia velutipes*. Molecular cloning and its overexpression to confer resistance to fungal infection in transgenic tobacco and tomato. *J. Biol. Chem.* **275,** 7230–7238 (2000).
52. Lauge, R. *et al.* Successful search for a resistance gene in tomato targeted against a virulence factor of a fungal pathogen. *Proc. Natl Acad. Sci. USA* **95,** 9014–9018 (1998).
53. Gilchrist, D. G. *et al.* in *Biology of Plant-Microbe Interactions* (eds de Wit, P. J. M., Bisseling, T. & Stiekema, W. J.) 406–410 (Int. Soc. Mol. Plant-Microbe Interact., St Paul, MN, 2000).
54. Dickman, M. B. Trans-species transfer of apoptotic genes and transgenic plants developed thereby. Patent application WO0026391 <http://ep.espacenet.com> (2000).

Acknowledgements

The authors thank all colleagues in Syngenta and especially the team at Syngenta-MOGEN for helpful discussions and suggestions, and B. van Wezenbeek for critically reading the manuscript. It is impossible to include all approaches in a complex field such as this, and we apologize to all colleagues whose work was not referred to.

© 2001 Macmillan Magazines Ltd

Lab automation

Automation by lab robots, PC control and integration.

MiniTrak

Packard Bioscience
www.packardinstrument.com

Flexible robotic integration

The MiniTrak robotic liquid handling system is compatible with most SBS microplates, including 96-, 384-, 864-, 1536-, deep well and PCR formats. A MultiPosition Dispense (MPD) module combines liquid dispense, dispense head movement, and tip and plate handling into a single unit for easy automation control. The dispense head can travel in the *x*, *y*, and *z* axes, enabling a single dispense head to access up to nine configurable plate or reagent positions on the labware deck behind the bi-directional conveyor. Tip Load can be performed on the conveyor path or deck. The MPD series of pipettors (10- and 20-inch models) expands the functionality of the MiniTrak platforms as an integrated system or workstation for use in high-throughput screening, genomics, and clinical and diagnostic research. The MiniTrak is compatible with other robotic accessories including Packard's SideTrak pick and place robotic arm for integration of microplate devices, including readers, incubators, plate sealers and other devices.

Twister II

Zymark www.zymark.com

Automated microplate handling

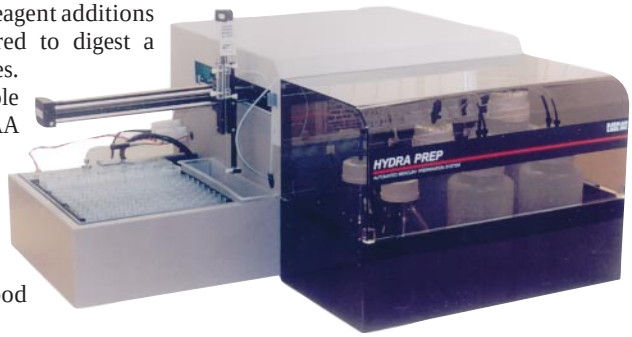
This latest incarnation of the Twister system features increased capacity, faster transport, and the ability to access the *x*–*y* index positions of almost any liquid handling deck. This low-cost microplate handler is designed to act as the hub for a new breed of 'Mini-Systems', providing users with an alternative to stackers and the larger, more complex integrated systems. Twister II's telescoping arm and rotating wrist joint make the system the only microplate handler that is compatible with virtually all microplate instruments. The flexibility allowed by the system's compact footprint and removable stacks make it possible to quickly and easily expand or reduce capacity, so the system never takes up more valuable laboratory space than is absolutely necessary.

Hydra Prep

Leeman Labs www.leemanlabs.com

Automated sample preparation for mercury analysis

Hydra Prep automatically digests samples according to operator selectable EPA methods. Once the protocol has been selected, Hydra Prep performs all reagent additions and heating steps required to digest a broad range of sample types. Hydra Prep is compatible with both the Hydra AA (cold vapour atomic absorption-based) and Hydra AF (cold vapour fluorescence-based) analysers in applications such as water, soil, fish, blood and tissue analysis.



Mercurial: Hydra Prep from Leeman Labs.

Span-8

Beckman Coulter www.beckmancoulter.com

Adding flexibility to automated liquid handling

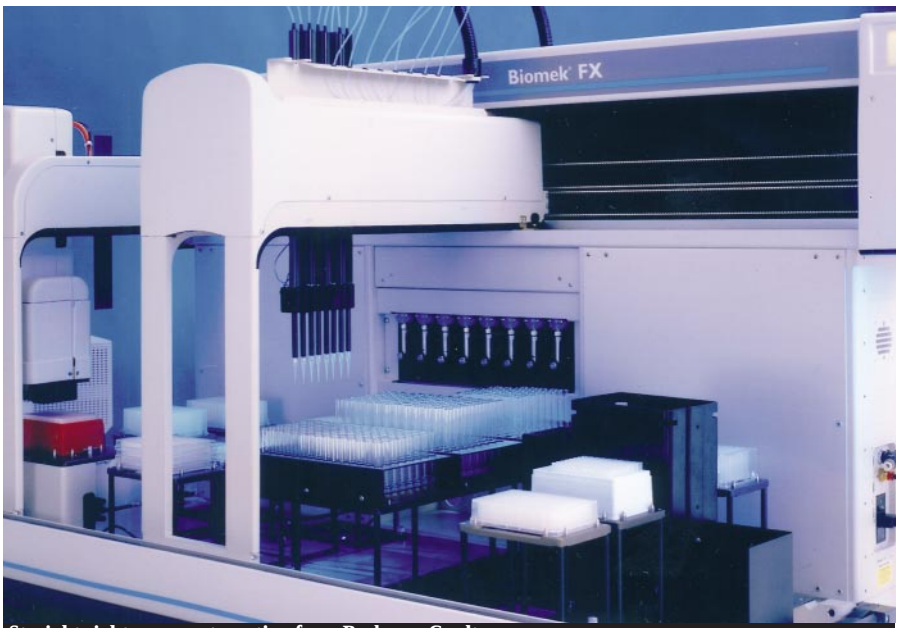
Span-8 is an independent eight-channel pipettor for the Biomek FX Liquid Handling System. The span range is 9–20 mm, making it suitable for both tube and microplate applications. Each of the eight pipettor channels has independent volume control, up and down motion and liquid level sensing. The pipettor uses fixed or disposable tips, or both, and tips can be easily interchanged. The ability to access single wells and tubes enables automation of applications such as serial dilution and hit picking for secondary screening. Available with single or dual pipetting arms, the Biomek FX Liquid Handling System allows the use of one or two pipettors. The Span-8 can be used in a stand-alone configuration or integrated with either a 96-channel head or the new 384-channel head to allow transfers from any format to any format, from tubes all the way up to 1536-well plates.

BioRobot 3000

Qiagen Instruments
www.qiageninstruments.ch

Keep at arm's length

Extended-arm systems have been added to the BioRobot 3000 so that high-throughput, walk-away sample processing can include multiple instruments outside the workstation. The extended arm can be used in either right- or left-arm configurations, allowing labware and accessories to be moved between the BioRobot 3000 worktable and a variety of external instruments. This enables full automation of protocols that include processing or analysis by instruments such as thermal cyclers or spectropho-



Straight eight: more automation from Beckman Coulter.

© 2001 Macmillan Magazines Ltd

tometers. The QIAsoft Operating System allows the extended-arm system to be easily integrated into existing laboratory environments.

RapidVap

Labconco www.labconco.com

PC-controlled evaporation

The RS-232 Link, available with RapidVap Vacuum and N_2 Evaporation Systems, makes it possible for operation functions to be remotely controlled from a personal computer. Functions such as preheat, vortex speed, temperature, time and vacuum or number of active nitrogen positions may be controlled through the link, and its two-way function means that the RapidVap can be controlled externally as well as providing status information for monitoring and recording parameters. Communications are in standard ASCII format. RapidVap Evaporation Systems use vortex motion, heat and vacuum or nitrogen blow down to speed evaporation of multiple sample volumes from a few millilitres to 450 millilitres.

1100 Series Autosampler

Agilent Technologies www.agilent.com

A new micro well-plate autosampler

This newly introduced module extends the capability of the Agilent 1100 Series capillary LC system. The Agilent 1100 Series micro well-plate autosampler adds higher flexibility and fast injection cycles to the system. This new module offers injection volumes ranging from nanolitres to microlitres, that are especially suitable for capillary LC. The Agilent 1100 Series micro well-plate autosampler can be operated through two controllers: the Agilent ChemStation or a handheld controller. Both provide simple-to-use graphical user interfaces for intuitive operation. Other accessories available include 96-well plates, 384-well plates, 96-deep well plates and cover mats.

CellScreen

Pierce www.piercenet.com

Coated microplates for cell-based assays including HTS and robotics

Coatings available for CellScreen microplates include poly-D-lysine, collagen I and poly-L-lysine in 96- and 384-well formats and in various colours. Following the coating procedure, all CellScreen Plates are treated with a predetermined irradiation dose to ensure consistency and suitability for cell-based assays. Applications include reporter assays, receptor binding assays and cytotoxicity assays.

AIS LIMS 7

Analytical Information Systems www.ais-lims.com

Out on a LIMS

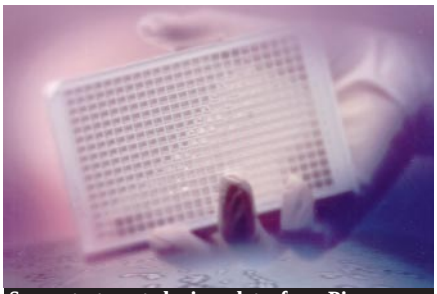
An integrated analytical quality control (AQC) module is at the heart of the latest release of Analytical Information Systems (AIS) Limited's Laboratory Information Management System. AIS LIMS Version 7 instantly alerts users when standards are required and warns if target values are not met. Another innovation is the integrated graphical wizard. AIS claims that while other LIMS systems require users to purchase third-party add-on packages, Version 7 saves money and time by carrying out its own graphical analyses. Version 7 also supports integrated third-party functionality and generates output compatible with Microsoft Office documents.

Series 200 Autosampler

PerkinElmer Instruments www.perkinelmerinstruments.co.uk

Productivity boost for HPLC applications

The Series 200 HPLC Autosampler can hold 225 samples, with up to three injections per minute. Users can obtain maximum precision from microbore to semi-prep injections using fixed or variable-fill modes. Pre-column dilution and derivatization are standard. With the addition of a Rheodyne 9725 PEEK-based valve, the Series 200 Autosampler may be used for biomolecular applications. The instrument may be used as a stand-alone unit, or may be controlled as part of a complete system, using the Turbochrom Professional Chromatography workstation. The Series 200 AutoSampler also offers complete procedures for IQ/OQ protocols, in order to comply with GLP guidelines.



Screen test: coated microplates from Pierce.

AlphaArray

Alpha Innotech www.alphainnotech.com

High-throughput microarray reader

The AlphaArray high-throughput reader is capable of accommodating a variety of slide- or plate-based array formats. AlphaArray is available in two different models: AlphaArray 7000, for automated multiple slide batch operation with eight excitation wavelengths and AlphaArray 6000 for single slide operation with two excitation wavelengths. Supplied and supported as an integrated unit, AlphaArray includes a high-performance Windows 2000/NT processor, reader unit, sample tray holder, and multiple copies of ArrayEase imaging software.

These notes are compiled in the Nature *office from information provided by the manufacturers.*

ADVERTISEMENTS

DIVERSE SMALL MOLECULES READY FOR SCREENING

CHEM BRIDGE™

10K–50K Compounds in Plates

· Next Day Delivery*
· Competitively Priced
*(Some Restrictions Apply)

Chembridge Corporation • 800.964.6143
858.451.7400 • 858.451.7401FAX • E-mail: chem@chembridge.com

www.chembridge.com

Polyclonal Antibodies & Custom Peptides

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

ResGen™
Invitrogen Corporation

U.S. or Canada 800-533-4363
FAX 256-536-9016

Check out our web site at www.resgen.com

OAKTOWN PUBLICATIONS
100 Western Avenue, London W3 7TX
Tel: 020 8749 4555; Fax: 020 8740 7782

Two Seminal books by **Shafi U. Ahmed**
Fermat's Last Theorem: New Simple Proof
(also Goldbach), ISBN 094 8331 100, and
We Are Alone: The Limited Universe
(Proves speed of light is immeasurable)
ISBN 094 8331 151. Price £10 each plus
P&P £1(UK), £3(abroad).
Order through Booksellers or from us. Major Credit/Debit Cards accepted. Key texts free on:
www.oaktownpublications.com

© 2001 Macmillan Magazines Ltd

More and more scientists are finding stimulation, fulfilment and pecuniary rewards in the financial sector, says Paul Smaglik.

Ian Smith was intrigued enough by the intellectual allures of the stock market that he abandoned basic science. Michael Steinmetz noticed that his pharmaceutical employer was spending less on research, so he left to launch a fund that would support more. David Shaw decided he could make a better living in investment than research.

Christopher Evans travelled from a postdoc post in Canada to stints as a UK biotech pioneer, to head of a biotech venture-capital firm. Roger Wyse travelled from a government scientist position to dean of two life science departments, to investment bank 'rainmaker'. Jesper Zeuthen shuttled between academic and industrial labs while consulting for a bank before finally joining the bank full-time.

These scientist–financiers found as many reasons to leave research careers as they discovered routes to depart. But their collective tales reveal shared themes as well. All had little or no formal finance training. All say that the intellectual joys of analysing a variety of companies, affixing values on them and deciding whether or not to invest in them outweighs being narrowly focused on one scientific project, or being weighed down with the management tasks that inevitably burden a mid-career scientist in both academia and industry. And all would admit that the money is not bad either.

Managing director Ian Smith finds finance more stimulating than bench work.

Moves into finance from both academic and industrial research are not just becoming more common, they are becoming more accepted in the scientific community. That was not always the case. Smith, managing director of Lehman Brothers Pharmaceuticals Research, considered by many to be the first basic researcher to make the switch from bench to bank in the mid-1980s, recalls that some colleagues perceived his decision to give up a basic-research position at SmithKline Beecham (now GlaxoSmithKline) to join Lehman Bros in London as 'traitorous'. Now scientists generally hold a "more enlightened" attitude about alternative careers in general, and towards finance in particular, he says.

He finds his finance position more stimulating than bench work. "The biggest intellectual challenge is placing value on companies — particularly loss-making biotech companies — and predicting the likelihood of their products going to the market," he says. He also likes the variety. "In many ways, it's a much more exciting environment than being in the pharmaceutical industry, because you get to see a diversity of scientific problems." He recommends that people considering a career in life-sciences finance spend some time in biotech or pharma first.

LEARN AS YOU GO

That advice holds true for Steinmetz, a general partner with MPM Capital in Cambridge, Massachusetts, who gradually attained business acumen while advancing his scientific career. After a biochemistry postdoc at the California Institute of Technology, the European native joined the Basel Institute for Immunology. While there, he received funding from Roche and got to know the company's head of global research, Jürgen Drews, who convinced Steinmetz that joining Roche would allow him to have broader scientific interests and to translate his basic work into practical applications.

When the company acquired Genentech in 1990, Steinmetz moved to New Jersey to run preclinical research and development for Roche in the United States. He was also responsible for Roche's global biotech efforts and forged several deals, among them Roche's collaboration with Millennium, where he joined the board of directors. But Roche, like other pharmaceutical companies going through tumultuous periods of mergers and acquisitions, began to reduce and restructure more and more of its drug research. So Steinmetz decided to use his long-held scientific background to evaluate his newer area of interest, biotech companies.

He says that it is easier for a scientist to learn finance than for someone with standard business training to learn science. He has used his knowledge to help set up two biotech-heavy funds — both worth hundreds of millions of dollars. He calls his work 'exciting' because it puts him in touch with scores of

scientists and start-up companies trying to redefine the cutting edge. But he cautions that there are probably not hundreds of full-time finance positions available for PhD scientists now. About 15 professionals at MPM manage $800 million in investments.

Shaw, president of DE Shaw, a private New York investment bank, agrees that opportunities are limited at the moment; he estimates that the firm receives about 400 applications for every opening. However, as technology becomes more complicated, he expects larger investment banks to hire more scientists, engineers and mathematicians.

Shaw did his PhD at Stanford in artificial-intelligence research and served on the faculty at Columbia University but realized when he wanted to start a family that he would need to make more money if he wanted to remain in New York, so he joined Morgan Stanley as an analyst. However, his academic bent kept him from remaining there long term. He was curious to see if his theoretical musings had any real-world application. "There was a tremendous intellectual challenge, as it was widely regarded to be impossible to use mathematical techniques to beat the market systematically," he says.

He started DE Shaw, which he calls a 'research enterprise' in part to find the answer. The company's success in using maths and statistics to look for stock-market opportunities has allowed the company to carry that label legitimately. In carrying out both financial and technological activities, he hires a "disproportionate" number of PhDs. As a result, the firm differs from more buttoned-down Wall Street houses. "The place doesn't look like an investment bank," he says.

Evans is all for bucking tradition. He, like Shaw, is unapologetic for opting out of academia to make money. One of the benefits of success, he says, is driving his children around the country in a Rolls Royce, blasting Black Sabbath. He also enjoys using his scientific background to make better investment decisions and his investment successes to support new scientific endeavours. After launching several companies, including Enzymatix and Chiroscience, the Welshman started Merlin Ventures, a fund that supports a broad array of biotech.

That company employs about seven or eight PhDs to analyse the science the company funds. Evans, like Shaw, believes that scientists do a better job of evaluating biotech than someone with traditional business training. Evans predicts that as multidisciplinary science begins to bring products to market, there will be a need for scientists with more diverse backgrounds to evaluate them.

PROS AND CONS

Evans finds his company gratifying because the companies it funds may bring therapeutics to people who need them — Merlin-funded companies have about 110 medicines in trials. The venture also provides jobs for scientists.

Wyse, the managing director of Burrill and Co., a private merchant bank, says that in function if not output his last academic position — dean of the University of Wisconsin's College of Agriculture and Life Sciences — does not differ dramatically from his new one. "Deans spend a lot of time raising money," he says. "We spend a lot of time raising money. Deans invest in good, bright people. We invest in good, bright people." However, there are some differences. "The bottom line here is you've got to make money," he explains. In academia, new thoughts and excellent science count for more. In investing, good but not great science that has wide commercial promise trumps stellar science with esoteric applications.

Michael Steinmetz has used his scientific knowledge to broker biotech funds worth hundreds of millions of dollars.

He notes that scientists at different stages of their career are of different value to investment firms and, consequently, require different levels of additional training if they expect to be hired. Newly minted PhDs — who, he says, the firm does consider — may be well served in getting some finance training, or starting out as analysts with a smaller firm.

Zeuthen, managing director of BankInvest in Copenhagen, slipped gradually, almost seamlessly, into investment banking. After a PhD in molecular biology from the University of Copenhagen and a postdoc at Stockholm University, he joined Novo Nordisk, because academic research in Denmark was poorly funded at the time.

In time, he found himself doing less science and more management. After seven years, he happily responded to an offer to start up a new lab for the Danish Cancer Society. Within a few months of taking up that post, he agreed to start consulting for BankInvest, which was considering starting up a fund based on the biotech boom in the United States.

As the years passed at the Danish Cancer Society, he again found himself doing less science and more management. So last year, he was tempted when he received an offer to join BankInvest full time. When he consulted for the bank, he handled the technical evaluations of prospective investments, while others looked at companies' financial strengths. But after years of consulting, he found he had learned a lot of the business nuances. Now he's happy he made the switch. "What I like about the venture business is the hands-on opportunity," Zeuthen says. "It's more hands-on than being a middle manager."

David Shaw uses maths and statistics to look for stock-market opportunities.

The position does have its downsides, however; investment is much more intense. He's not sure he could hold this position as long as he could a university one. "If I had stayed in academia, I could have continued like that until I reached 90," he says. But he's not sure he would have wanted to. ■

Paul Smaglik is Naturejobs editor.